Patterns of Short-term Genome Evolution in E.coli and Shigellae

Balbi, Kevin

Award date:
2009

Awarding institution:
University of Bath

PATTERNS OF SHORT-TERM GENOME EVOLUTION IN *E. COLI* AND *SHIGELLAE*

Kevin Jon Balbi

A thesis submitted for the degree of Doctor of Philosophy

University of Bath
Department of Biology and Biochemistry

July 2009

COPYRIGHT

……

For my Grandad

“ $$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$ ”

Acknowledgements

I would like to thank my Supervisor Dr. Edward Feil for his guidance, help and instruction throughout my PhD. Additionally I thank Dr Eduardo Rocha for the genome alignments and for the software used in the initial stages of the analysis.

I'd also like to thank the members of the office/lab, both past and present, for their help and entertaining conversations, with a special thanks to Caroline, Ciara, Matt, Nicola, Steph and Pete whose friendship and company has made my PhD a thoroughly enjoyable experience. I must also thank my friend Sam for his assistance with matters mathematical over many a lunch and coffee break.

Ste, I thank for his tireless and unwavering encouragement which has kept me motivated and for his unerring ability to find the right thing to do or say to cheer me up when I've been feeling down.

Finally I'd like to thank my brother Ian and my parents Sue and Jon, for their love encouragement and support throughout my entire education.

Abstract

The time-dependence of molecular evolution, specifically over short timescales, has been shown to be a major confounding factor in the analysis of nucleotide changes between closely related strains or species. The assumption that selection works extremely quickly to purge all of the deleterious changes is at odds with the Nearly Neutral model of evolution, whereby the majority of changes are only mildly deleterious and therefore impose only a minor fitness cost so they are relatively rapidly purged only in populations with large effective population sizes.

The aim of this project was to explore the patterns of nucleotide changes evident between the core genomes of nine *E. coli* and *Shigella* strains, with the latter having adopted a specific ecological niche in the recent evolutionary past. The *Shigellae* and *E. coli* show little difference in their extant genome compositions, in terms of nucleotide composition and genome size, however there are a markedly higher number of pseudogenes and insertion sequences present in the *Shigella* genomes.

The polymorphism profiles of the core genomes reveal a time-dependency of d*N*/d*S*, Ti/Tv, +AT/+GC and the Metabolic cost of Amino acid changes, the nucleotide data showing a clear separation of the *E. coli* from the *Shigellae*, with the latter showing trends indicative of weaker purifying selection. Additionally these differences are evident when examining the nucleotide ratios (+AT/+GC & Ti/Tv) along the core genome, also revealing patterns of evolution associated with genome position. A simulation based approach reveals different projected nucleotide contents for the *E. coli* and *Shigellae* genomes further highlighting their different evolutionary paths as evident from the polymorphism profiles.

The methods employed and developed in this study provide a useful and effective toolset for examining the evolution of bacterial genomes over short timescales, especially in light of the availability of multiple whole genome sequences for a given ‘species’.

Table of Contents

Abbreviations

Abbreviation	Definition
~P	High energy Phosphate bond, as carried by ATP or GTP
BASH	**B**orne-**A**gain **SH**ell
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
DNA	**D**eoxyribo**N**ucleic **A**cid
EIEC	**E**ntero**I**nvasive **E**scherichia **c**oli
Gbp	Billion nucleotide base pairs
GTR	**G**eneral **T**ime **R**eversible (referring to a substitution model)
IS	**I**nsertion **S**equence
Kbp	Thousand nucleotide base pairs
Mbp	Million nucleotide base pairs
MK Test	**M**acDonald-**K**reitman Test
ML	**M**aximum **L**ikelihood (referring to the method of phylogenetic inference)
MLEE	**M**ulti**l**ocus **E**nzyme **E**lectrophoresis
MLSA	**M**ulti**l**ocus **S**equence **A**nalysis
MLST	**M**ulti**l**ocus **S**equence **T**yping
MSSA	**M**ethicillin **S**usceptible **S**taphylococcus **a**ureus
NCBI	**N**ational **C**entre for **B**iotechnology **I**nformation
N-J	**N**eighbour-**J**oining (specifically referring to the method of inferring phylogenetic trees)
NQ Sites	Non-fourfold degenerate / **N**on-**Q**uartet nucleotide sites
ORF	**O**pen **R**eading **F**rame
PC	**P**rincipal **C**omponent
PCA	**P**rincipal **C**omponent **A**nalysis
Q Sites	Fourfold degenerate / **Q**uartet nucleotide sites
SNP	**S**ingle **N**ucleotide **P**olymorphism
Spp	Species
TBS	**T**erminal **B**ranch **S**ubtraction
TTSS	**T**ype **T**hree **S**ecretion **S**ystem
VP	**V**irulence **P**lasmid

Figure List

Equation List

Table List

Chapter 1 - Introduction

1.1 – In the Beginning…

1.1.1 – Natural Selection

In 1837 the ornithologist John Gould, having been studying the birds that Charles Darwin had brought back from the Galapagos Islands, commented to Darwin that what had been believed to be an assortment of blackbirds, gros beaks and finches where in fact a collection of 12 species of finch. Upon re-examination of his notes and those notes of the other crew of H.M.S. Beagle Darwin soon realised that each separate species of finch came from a distinct island.

During 1838 Darwin read Malthus's "An Essay on the Principal of Population" which, in concert with the struggle for existence he had observed during his studies of plants and animals, caused him to speculate that;

> *"favourable variations would tend to be preserved and unfavourable ones be destroyed. The results of this would be the formation of new species.*

[Charles & Francis Darwin, The Life and Letters of Charles Darwin, p83]

In 1855 a younger naturalist, Alfred Russel Wallace, at the time making his own studies of the nature of species, via fieldwork in the Malay Archipelago, noted that *"every species has come into existence coincident in space and time with a closely allied species"* an observation published a paper which was brought to Darwin's attention in 1856 by his colleague, Charles Lyall who in turn prompted Darwin to publish his theories to establish priority.

However in early 1858, only halfway through his 'big book on species' Darwin received an essay from the young Wallace entitled "On the tendency of Varieties to Depart Indefinitely from the Original Type". At the suggestion of Lyall and others, extracts from Darwin's work and the work of Wallace were read at the Linnaean Society on the 1st of July 1858. Where they were seen as either insignificant or unworthy of consideration.

It wasn't until the publication of Darwin's magnum opus "On the Origin of Species by Means of Natural Selection" on the 22nd November 1859 that the significance of both Wallace and Darwin's earlier works became apparent. In the introduction of which Darwin very succinctly lays out what both he and Wallace had come to believe with regards to the organisation of life and the relationship between species;

> *"As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form."*

[Charles Darwin, On the Origin of Species, p5]

In that single section of his introduction he very succinctly laid out the entire premise of his theory of "Natural Selection".

1.1.2 – The problem of Inheritance

Whilst Darwin and Wallace had defined a role for adaptation and consequently a plausible system for change within species which also provided an explanation of the rise of new species, they had yet to ascribe a model for the inheritance of traits and characteristics. The prevailing view, originally proposed by Lamarck, of the "soft inheritance" model – which permitted the inheritance of acquired traits and characteristics, was thought to be plausible and Darwin considered it a possible source of the change on which natural selection acts.

A German biologist by the name of Friedrich Weissmann originally believed, as Darwin, in this model of inheritance. However during his work on embryology in Sea Urchins noted that there were two kinds of cellular division, which he coined "equatorial" and "reductional", the latter he concluded resulted in the generation of Germ cells, which would carry hereditary information to the next generation. Whilst his assumptions on this were based on his now defunct "Germ Plasm" theory, it did prompt the rediscovery of the work of Gregor Mendel.

One of the earlier popularisers in the support of the rediscovered Mendelian model of inheritance was William Bateson, he championed Mendel's "Genetic" system of inheritance over the prevalent focus on statistical analysis of phenotype variation. The two laws proposed by Mendel were;

1. **Law of Segregation:** Each pair of alleles (all alleles occur in pairs) separates in the formation of gametes.

2. **Law of Independent Assortment:** Two or more alleles segregate independently during gamete formation.

However it was believed by many than this model of inheritance wasn't sufficient to explain the maintenance of diversity in the population, many arguing that dominant alleles would increase in the population.

1.1.3 – Population Genetics

The problem of proving that Mendelian law could be applied to whole populations and still result in stable diverse populations, which have not become homozygous for the dominant allele, was introduced to Godfrey Hardy (an English mathematician). Hardy wrote a short note to the editor of Science in 1908 illustrating through *"very simple"* maths (Hardy 1908) that the ratio of alleles from any given generation to the next will remain unchanged, assuming the population size is large enough that mating is random and that there is no inherent bias in the distribution of alleles between the sexes. This same principal having been formulated independently by Wilhelm Weinberg also in 1908 became known as the Hardy-Weinberg principle, and states that, barring the introduction of an outside influence the allele and genotype frequencies within a population remain stable.

During the 1920's & 30's the work of R. A. Fisher, J. B. S. Haldane & Sewall Wright formed the basis of population genetics, building on the work done by Hardy & Weinberg. Fisher showed in a series of papers how the action of several discrete loci can produce the continuous variation (as opposed to discrete) observed by biometricians and that Mendelian genetics was wholly compatible with evolution driven by natural selection.

Haldane applied mathematics to real world populations adding further weight to the observations and conclusions of Fisher and Wright was responsible for the introduction of the idea of an adaptive landscape, suggesting that cross breeding and genetic drift, as exhibited in small populations, could drive populations away from an adaptive peak, creating the potential for natural selection to push them towards new and different adaptive peaks.

1.1.4 – The Modern Synthesis

The “Modern Synthesis” was a convergence of ideas that brought together naturalist, genetic and paleontological evolutionary studies over the course of the 1930’s and 40’s. It provided a clear and coherent model of evolution, as a process that occurs gradually via small genetic changes, where natural selection is the primary mechanism of change and factors such as genetic diversity in a population, the population size and ecological niche are all important. The differences between species were considered to have occurred gradually through processes such as geographical separation and extinction.

This establishment of a clear model provided the impetus to investigate evolutionary trends and phenomena into the mid 20th century.

1.2 – The Molecular Era

1.2.1 – Molecular Genetics

During the late 1930's and early 1940's George Beadle and Edward Tatum were studying the interplay between genes and metabolic processes, in contrast to the trend at the time of working out the genetic basis of a given phenotype. Using X-rays they performed mutagenesis experiments in various species in the fungal genus of *Neurospora*, in which they demonstrated that mutation events resulted in a specific metabolic deficiency, the inability to synthesise vitamin B_6 for example. It is mentioned at the end of the paper in a note that these deficiencies were inherited as though they were *"differentiated from normal by single genes"* (Beadle and Tatum 1941). Establishment of this one to one link between enzymes and genes would prove to be a key realisation in establishing the molecular basis of inheritance and the storage material of genetic information.

Work in 1943 on the transformation of type II *Pneumococci* to type III, via the introduction (under specific conditions) of a highly purified fraction isolated from heat-killed type III cells, strongly hinted that Deoxyribonucleic Acid had a role to play in heredity and genetics (Avery, MacLeod et al. 1995, originally published 1944). This was then confirmed in 1952 when, via radio-labelling studies of protein coat and DNA, DNA was proven to be the genetic material of the T2 phage (Hershey and Chase 1952). The following year, based upon X-ray crystallography performed by contemporaries (Franklin and Gosling 1953; Wilkins, Stokes et al. 1953) and using information determined by Erwin Chargaff on the relative proportion of each of the four bases known to be present in DNA (Chargaff 1950), James Watson and Francis Crick proposed a viable structure for DNA (Watson and Crick 1953) and in a short paragraph towards the end of the paper they suggest that *"…the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"*. This put genetic and evolutionary studies on a molecular footing, providing the potential for more fundamental insights into both.

1.2.2 – Non-Darwinian Evolution

Estimation of the genome wide rates of amino acid change by Motoo Kimura in 1968, lead him to conclude that these rates would be equivalent to an average of 1.8 nucleotide changes per year for the human genome, assumed to be approximately $4x10^9$ bp at the time. Which in light of the "substitutional load" tolerable by mammalian genomes, according to JBS Haldane's work (Haldane 1957), and the prevailing view at the time, that the vast majority of substitutions were deleterious and a minority advantageous, would have been completely untenable within any organism. This prompted the supposition that many of these changes must be somehow invisible to selection, that is to say they must be 'Neutral' changes, conferring neither benefit nor hindrance to the organism. It was also shown that if the mutations were predominantly neutral, a high rate of mutation could be compatible with a low substitutional load (Kimura 1968).

Given the preponderance of neutral mutations, it becomes necessary to consider processes other than just selection as responsible for the shaping of genomes. Genetic drift would largely dominate the accumulation or loss of neutral mutations, with the proportion of the next generation inheriting a neutral mutation dictated mainly by random chance until, eventually, the mutation 'goes to ground' that is it either reaches fixation or is lost entirely.

A more rigorous exploration of this model agreed with Kimura that, a framework of evolution incorporating drift and selection, with a significant proportion of neutral mutations is more likely than one driven purely by selection (King and Jukes 1969). Including the observation that of the 549 possible single nucleotide changes from any given codon to any other, 134 are synonymous – they result in no amino acid change and are therefore completely invisible to selection. King and Jukes also illustrated there are certain amino acid changes, dependent upon position within the protein and function of the protein, that are effectively neutral having little or no impact on the protein's function.

Subsequently Tomoko Ohta, a student and colleague of Kimura, proposed a modification to the strictly neutral model. Observing that whilst there are truly neutral mutations, there

are others which may be only very slightly deleterious, or 'nearly neutral'; *"Any amino-acid substitution of a well organised molecule is likely to disturb that organisation"* (Ohta 1973). Even a 'conservative' amino-acid substitution preserving the general function of, for example, an enzyme may have an effect, albeit a small one. In small populations the chance of this slightly deleterious change increasing in the population or even reaching fixation through drift before it is purged by purifying selection, is higher than that in larger populations. That is when $N_e s > 1$ (where N_e is effective population size and s is the selective coefficient of the mutation) selection will dominate the fate of the mutation as there are sufficient members of the population to 'select' from or the mutation confers an immensely significant advantage or disadvantage that it is selected for or against with ease. Where this value is less than one, i.e. the mutation either has a small/insignificant effect or the effective population size is small then genetic drift will dominate the fate of the mutation as selection will take far longer to purge/accept the mutation than its loss or fixation by random genetic drift. A diagrammatic comparison of these models is shown below (figure 1.2.2a).

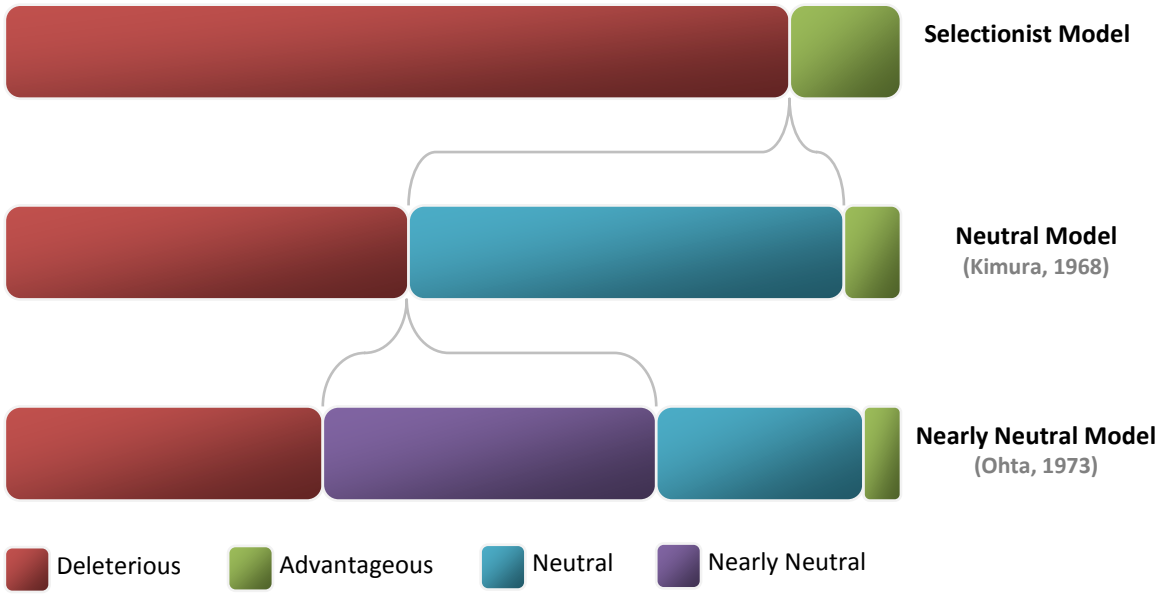


Figure 1.2.2a – A diagrammatic representation of the different models of molecular evolution. Bar sizes are representative only, not to scale.

1.2.3 – Time Dependence of Molecular Evolution

The d*N*/d*S* ratio has commonly been used as a convenient way to assay the evolutionary pressures acting upon genes, given that synonymous changes are largely regarded to be neutral or have a much smaller fitness impact than nonsynonymous changes. So the ratio has been considered to be able to reveal the type of selection at work when comparing two orthologous sequences. With values less than one, considered to reflect purifying or "stabilising" selection and values greater than one reflecting diversifying or "positive" selection. Whilst d*N*/d*S* is a convenient metric there are several other classes nucleotide changes or genomic features which can also indicate the selective pressures acting upon a genome, they are listed below (Table 1.2.3a).

Type of Deleterious Change	Effect	Reference
Increased GC → AT mutations	Weaker base pairing AT Rich Codons → Greater metabolic cost of encoded AA C→T most common mutation (deamination of methyl-C→U)	(Knight, Freeland et al. 2001) (Wernegreen and Funk 2004)
Increased proportion of Transversions	Transversions are proportionally more nonsynonymous Nonsynonymous transversions are less conservative AA changes than nonsynonymous transitions	(Freeland and Hurst 1998) (Zhang 2000)
Increased cost of amino acid changes	Increased AA cost → increased metabolic load on cell	(Swire 2007)
Increased number of pseudogenes	Non-functional, potential for reactivation	(Cole, Eiglmeier et al. 2001)
Increased number of insertion sequences	Genetically disruptive Unpredictable, inserting into potentially essential genes	(Dale and Moran 2006) (Escobar-Paramo, Ghosh et al. 2005)

Table 1.2.3a – A list of deleterious genomic changes used to indicate selective pressures, other than dN/dS.

It has been assumed that comparison of any two sequences would yield equally valid results with regards to evolutionary patterns, such as those inferred from the value of the d*N*/d*S* ratio (Kimura 1991). This holds in the vast majority of cases where there is a significant evolutionary distance or divergence time between the two sequences being compared, where last common ancestors are millions of years ago. This type of comparison has until recently been the predominant examination of evolutionary

pressures/trends in molecular data. However the completion of multiple genome sequences for many bacterial taxa or species has enabled statistically rigorous examination of more closely related sequences using the same techniques; the availability of large quantities of sequence data for close comparisons is a necessity given the relative paucity of nucleotide differences between members of the same species or genus than comparisons of species with millions of years of unshared evolutionary history. This paucity of differences eliminates complications associated with multiple hits, however it means that the relative consequence of sequencing error is far greater, however based upon the error rates seen in genomes sequenced at the Sanger Centre of 0.37 bases per genome (value taken from Rocha et al 2006, calculated from the MSSA476 genome and corresponding to 1 error per 7807752 bases) and analysis by other groups (Gutacker, Smoot et al. 2002; Read, Salzberg et al. 2002) showing that 90% or more of synonymous changes are accurate, the vast majority of differences can be assumed to be accurate.

Nonsynonymous differences observed at high levels of divergence are largely fixed differences, however in cases of comparison more closely related species, or even strains of the same species the assumption that the observed difference is fixed in the population is no longer valid. The view of the process of selection as a quick and efficient 'check and purge/accept' system is not compatible with the Nearly Neutral Theory nor is it compatible with recent observations (Feil, Cooper et al. 2003; Ho, Phillips et al. 2005). Feil et al noting in MLST data of 334 strains of *Staphylococcus aureus* that the proportion of nonsynonymous changes observed between two more closely related haplotypes was higher than more distant comparisons. Selection takes time to act upon a mutation once it has been incurred by an organism, with mutations of greater impact being acted upon more quickly than those of lesser or no impact. This 'time lag' between mutation and selection is not a fixed quantity, but rather is dictated by the same forces that were observed to govern the selection/drift balance (Rocha, Maynard Smith et al. 2006), that is the time taken for selection to 'process' a mutation is a product of the effective population size of the host organism (what selection has to 'work with') and the selective coefficient of the mutation (how much selection 'cares').

Various observations have been made that are consistent with this 'time lag', specifically observation of relatively higher proportions of nonsynonymous changes (which are more likely to be deleterious), versus synonymous changes (which are effectively neutral) between more closely related organisms (Jordan, Rogozin et al. 2002; Feil, Cooper et al. 2003; Ho and Larson 2006). In many cases it had been dismissed as either a statistical artefact or interpreted as evidence or positive or relaxed selection.

An examination of this trend through comparative genomics was carried out by Rocha et al (2006); their comparisons of strains/species of several genera of bacteria revealed a consistently higher proportion of nonsynonymous changes between more closely related species or strains, the relationship between two strains being measured using percent divergence in intergenic regions as it is independent of both d*N* and d*S* (figure 1.2.3a).

They observed a clear and striking difference in the trends observed between the *Escherichia* and the other taxa examined, a difference which agrees with their simulation analysis (figure 1.2.3b) and highlights the possibility that population size is governing the differences observed. The effective population size (N_e) of *E. coli*, estimated to be approximately 10^8 (Hartl, Moriyama et al. 1994), owing to it being an ecological generalist, is very likely to be higher than that of the other taxa examined as they have rather more restricted ecological niches.

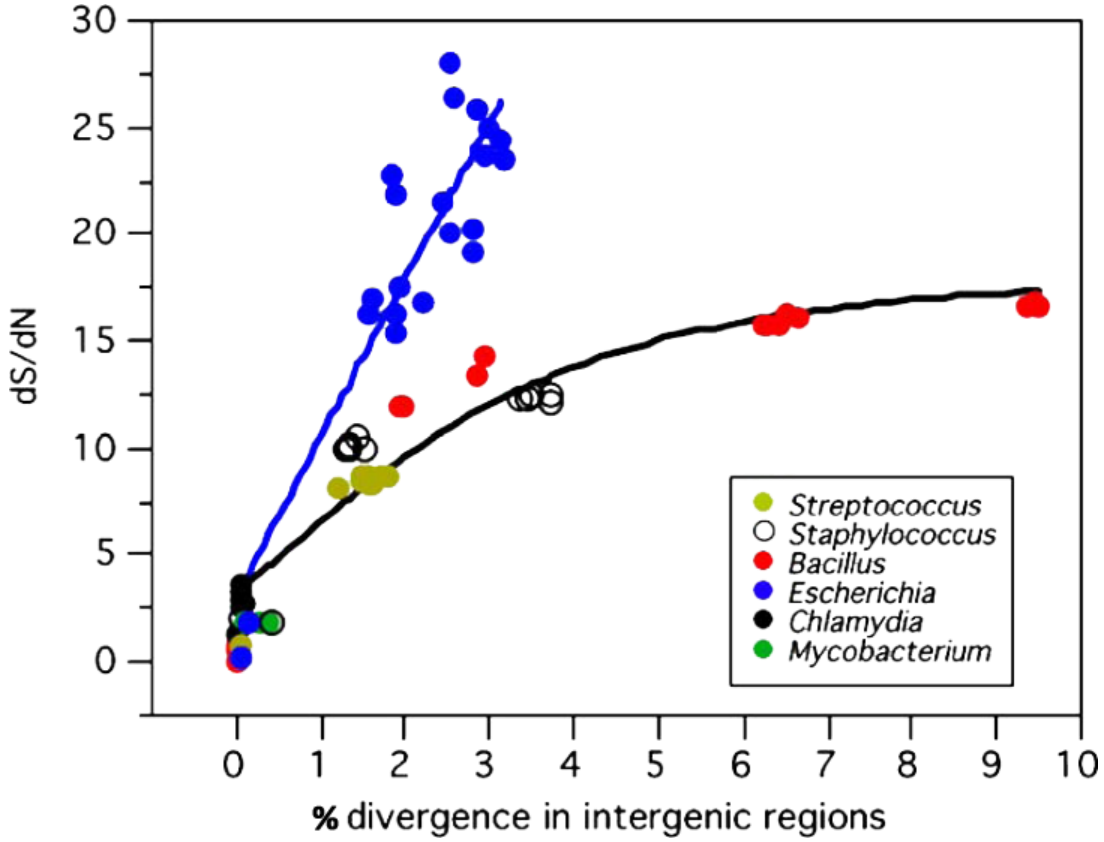


Figure 1.2.3a – The dS/dN ratio oberserved in comparisons of two members of the same species/genus against % distance in intergenic regions. Each point represents a single pair-wise comparison. Figure reproduced, with permission, (Rocha, Maynard Smith et al. 2006).

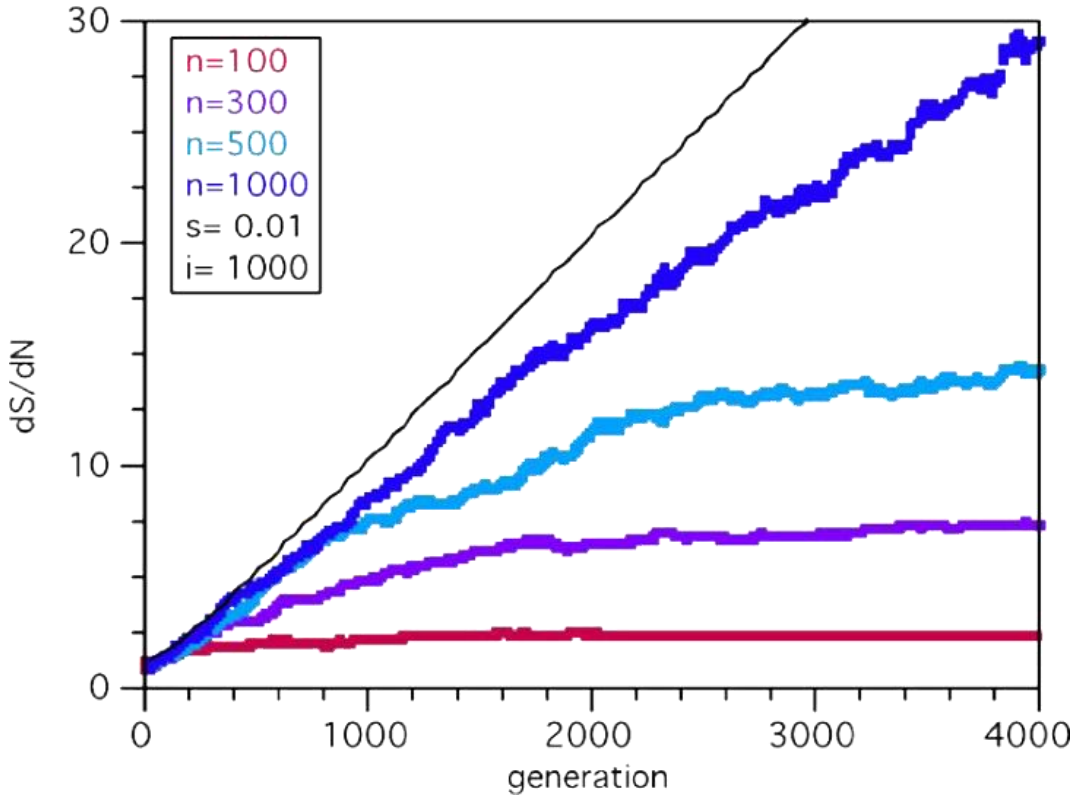


Figure 1.2.3b – The dS/dN ratio oberserved during simulations of populations with different effective population sizes. Each simulation was run 1000 times for 4000 generations, the black line represents an infinite effective population size. Figure reproduced, with permission, (Rocha, Maynard Smith et al. 2006).

Consequently it can be considered that an observation of genetic differences (nucleotide or amino acid) between more closely related sequences, two strains of the same species for example, predominantly reflects mutation biases (red lines in fig 1.2.3c), whereas a comparison of more divergent sequences, such as that between two different species in different genera, reflects more the bias imposed due to selective pressures (blue line in figure 1.2.3c).

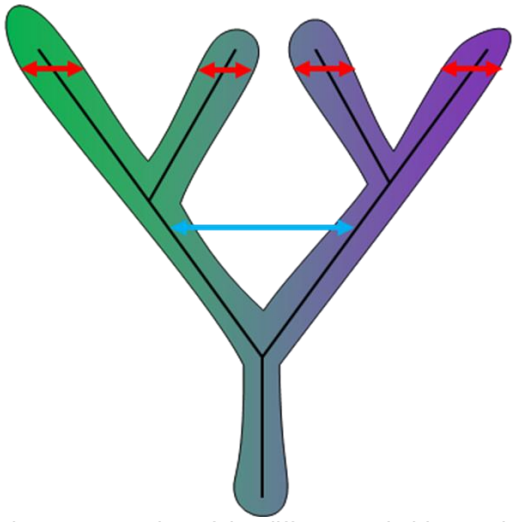


Figure 1.2.3c – A diagrammatic representation of the differences in biases observed when comparing sequences within species (red lines) and between species (blue lines), using a hypothetical tree containing two 'species' represented in Green and Purple.

One example of the potential pitfalls associated with neglecting to take this into account is a study by Jordan et al (Jordan, Kondrashov et al. 2005), where they analysed sets of three complete genomes, comprising two sister taxa and an outgroup, across a range of organisms representing the Bacteria, Archaea, Yeast and Mammals, in the case of the Hominidae, comparing *Homo sapiens* with *Pan troglodytes* and using *Mus musculus* as an outgroup. Polarising all observed amino-acid changes using the outgroup, Jordan et al observed an apparently universal bias in favour of removal of the 'older' amino-acids in favour of the 'newer' ones, and so concluding that the amino-acids more recently adopted for protein manufacture have not yet reached an equilibrium with the oldest ones, (see figure 1.2.3d).

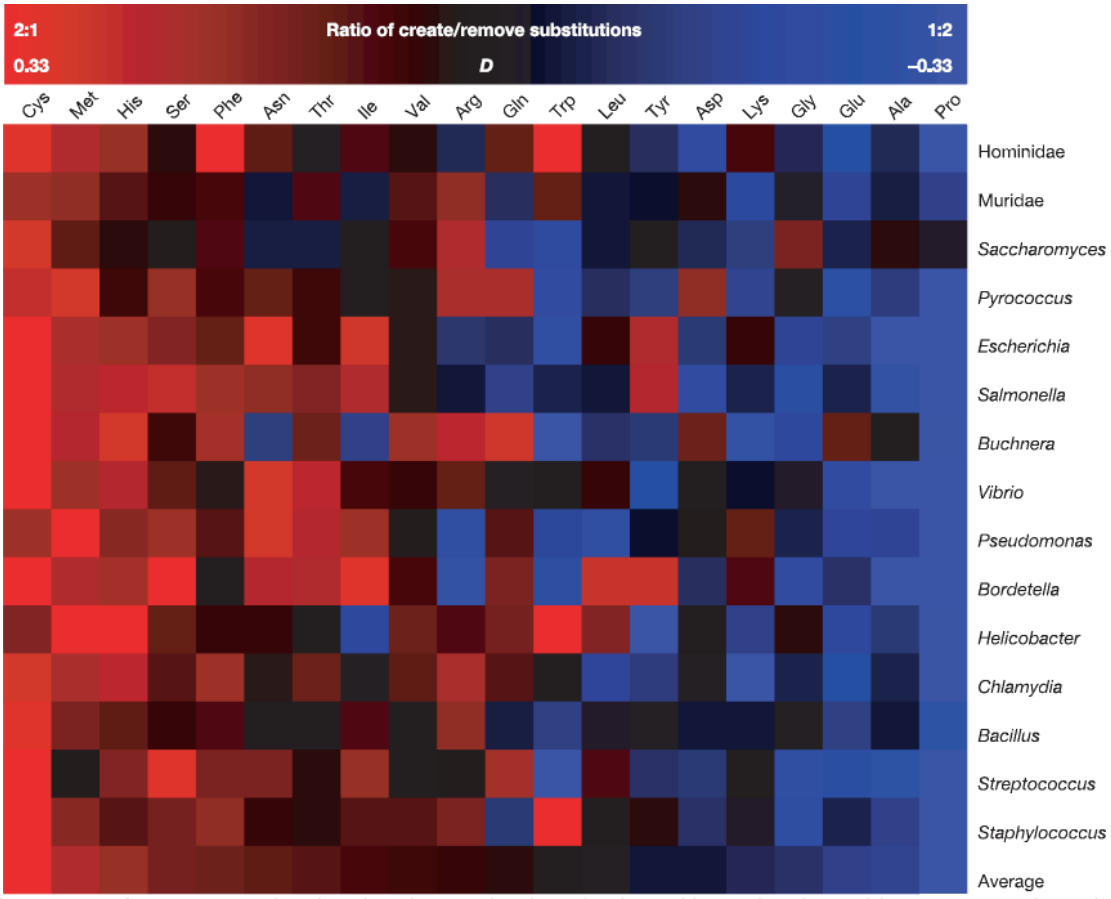


Figure 1.2.3d – Heatmap showing the observed ratios of gain and loss of amino acids across a variety of taxa, showing a hypothesised universal trend related to the order of recruitment of the amino acids to the genetic code. Adapted from Jordan et al (2005).

However, given these comparisons are only of closely related species or even strains, it can be more correctly concluded that the trend observe mainly reflects the mutation bias (Hurst, Feil et al. 2006), and so rather than there being a universal selection bias in favour of the 'newer' amino-acids there is the reverse – there is a mutation bias in their favour, suggesting that they have been selected against as mutation would favour the restoration of equilibrium against any selective bias, a trend that has previously been observed (McDonald 2006). Analysis of the cost per amino acid replacement, performed by Hurst et al, revealed a strong correlation with the log number of changes (figure 1.2.3e), supporting the notion that the observed trends reflect only the short term mutational bias rather than a longer term selective bias.

This situation makes sense from an energetic perspective; many of the newer amino-acids are manufactured by extending the metabolic processes associated with pre-existing ones, they are more costly to include into the proteins, meaning that where the function of the protein can be retained any mutation that reduces the overall cost

(replacing a ‘newer’ amino-acid with an ‘older’ one) would confer a small advantage to that organism. Hurst et al, have also demonstrated that the putative order of recruitment of amino acids suggested by Jordan et al correlates strongly with the metabolic cost of the amino acids (Akashi and Gojobori 2002).

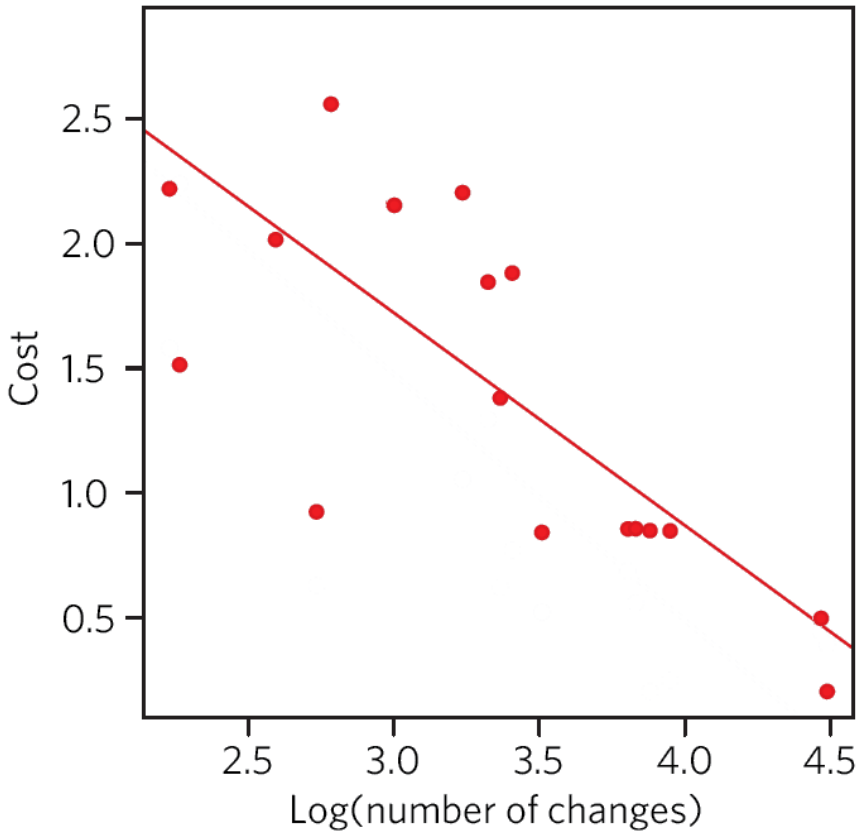


Figure 1.2.3e – Scatter plot of the average cost of amino-acid change (Y-axis) versus nonsynonymous distance between sequences (X-axis). The line indicates the regression $R^2 = 0.601$ $p < 0.0005$. Adapted from Hurst et al (2006).

Furthermore recent work by Hughes et al (2008) has shown that in the more common measures used to indicate the presence of positive Darwinian selection, this time-lag between mutation and purifying selection undermines the usual interpretation of these results. The NI value is a ratio of; the ratio of nonsynonymous to synonymous nucleotide diversity within species (π_A/π_S) to the ratio of nonsynonymous to synonymous substitutions between species (dN/dS or k_A/k_S), values of NI < 1 are interpreted as evidence of positive Darwinian selection. Hughes et al reported that, in their dataset of 12 bacterial species, a value of NI < 1 was more strongly correlated with a low value of π_A than a high value of k_A, suggesting that NI < 1 is more likely an indicator of effective purifying selection within species (lower π_A/π_S) than of positive Darwinian selection

between them (higher d*N*/d*S*). This has implications too for the interpretation of the MacDonald-Kreitman (MK) test, which itself relies upon NI value and the outcome of the G-test. Hughes et al show that the G-test within those genes with an NI value < 1 is more likely to be significant when there is an absence of rare nonsynonymous polymorphisms, which in turn reflects the action of purifying selection as these polymorphisms are often slightly deleterious (Hughes, Packer et al. 2003). Consequently the MK test will often succeed in highlighting genes as under strong positive selection when they are in fact experiencing strong purifying selection.

1.2.4 – Effects of Population Size

It is an intrinsic aspect/consequence of the Nearly Neutral Theory that the ability of the evolutionary process to clear a deleterious mutation is based on two different factors. The more intuitive of these being the severity of the effect that the mutation has, its 'selection coefficient' (*s*), it would stand to reason that mutations conferring a larger adaptive benefit or more likely a significantly maladaptive mutation would result in a selective response whereby the more or better adapted organisms outcompete the less well adapted. However in cases of only slightly deleterious 'Nearly Neutral' mutations the maladaptive cost may not be significant enough for selection to disfavour the organisms carrying it by any appreciable margin over those that do not. In that case stochastic survival of the mutated genotype in the population (genetic drift) becomes important, and with it the effective population size. In the case of a large effective population size the chances of a single slightly maladaptive mutation being randomly sampled are slim; however the probability of sampling the mutated genotype from the population increases as the reciprocal of the effective population size. In cases of very small effective population sizes the mutation may well reach fixation via genetic drift before it can be eliminated from the population by selective pressures. The figure below (figure 1.2.4a) illustrates this effect, a simulation of two populations one comprising 100 individuals (lower panel) the other 10 individuals (upper panel), in each case 20 unlinked alleles are represented and their abundance in the population tracked over 50 generations under a Fisher-Wright model of genetic drift. In the smaller population the vast majority of the alleles have been either fixed or lost after 50 generations, whereas in the larger population the converse is true – the vast majority of the alleles are still only present as part of the standing variation, thus giving selection more time to act.

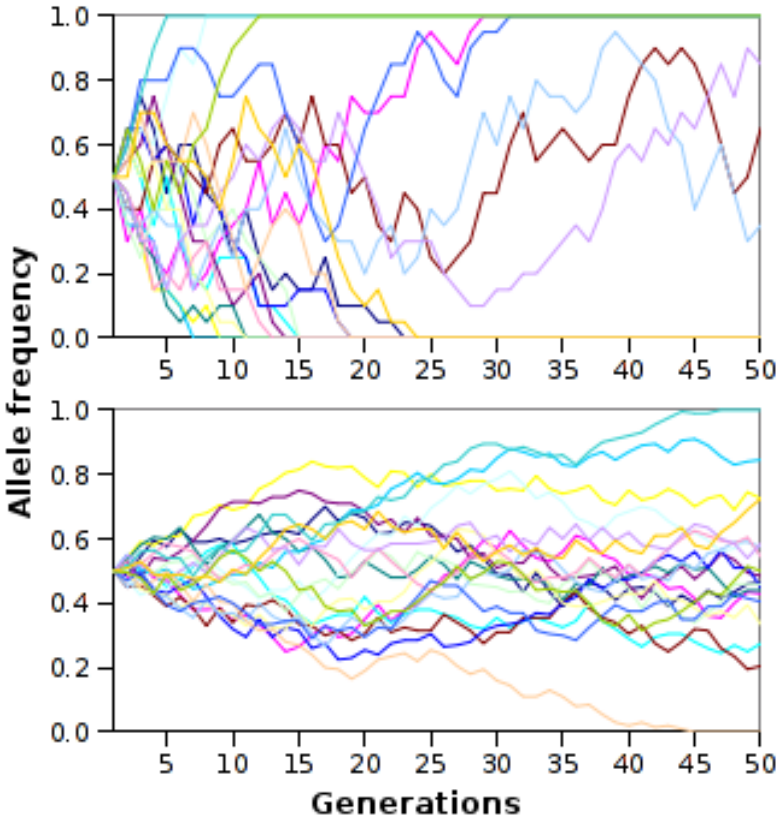


Figure 1.2.4a – Plots representing the simulated abundance of 20 unlinked alleles in two populations of different sizes, one of 10 individuals (upper panel) and one of 100 individuals (lower panel) under a Fisher-Wright model of genetic drift. Image taken from Wikipedia article: "Genetic Drift", accessed 12/01/2009. http://en.wikipedia.org/wiki/Genetic_drift

There are several examples of studies where observations congruent with this effect have been made; in birds, specifically species of doves and ducks, it was noted that populations living on islands, consequently having smaller effective population sizes and a fair degree of reproductive isolation from mainland groups, showed a higher number of non-synonymous changes in their mitochondrial genomes than equivalent populations living on the mainland, with no significant difference in synonymous changes (Johnson and Seger 2001). A similar study, sampling a wider range of organisms, noted the same surfeit of nonsynonymous over synonymous changes in island as compared to mainland populations in 70 phylogenetically independent comparisons encompassing taxa from the vertebrates, invertebrates and plants (Woolfit and Bromham 2005).

This 'island versus mainland' effect has also been observed in prokaryotes (Escobar-Paramo, Ghosh et al. 2005), and can be most dramatically observed in those bacteria which have developed intimate co-evolutionary associations with Eukaryotes. Host

association of this nature has evolved independently in many bacterial taxa, resulting in cases of mutually obligate endosymbiosis or intracellular parasitism. Thought to have originally evolved from free-living generalists, these host-associated bacteria have all undergone a massive reduction in effective population size as a direct result of this ecological shift, specifically the resultant niche restriction and ecological isolation. In the case of vertically transmitted obligate endosymbionts of insects (where host progeny are infected via the mother's ovaries), this effect is reinforced by a bottleneck at every host generation, such that N_e of the bacteria begins to approach that of its host (Birky, Maruyama et al. 1983).

Genome sequences recovered from endosymbionts reveal striking footprints of increased deleterious change owing to drift, as in the island bird populations. Genomes of the genus *Buchnera* (an obligate endosymbiont of aphids) show a markedly higher proportion of non-synonymous changes than their closest free-living 'mainland' relatives, for example *Escherichia coli* (Wernegreen and Moran 1999). Woolfit and Bromham (2003) presented an ambitious study on 13 phylogenetically independent comparisons of bacterial and fungal endosymbionts with their closest known free-living relatives, in which they confirmed an increase in substitution rate in the endosymbionts, again consistent with increased drift, although much of the data was limited to 16S gene sequences rather than whole genomes. Other features of endosymbiont genomes which are consistent with a reduction in population size include; extensive genome degradation, AT enrichment, loss of selective codon bias, and deleterious mutations affecting 16S rRNA stability (reviewed in Wernegreen 2002).

Hallmarks of increased genetic drift have also been observed in several pathogenic species (Andersson, Alsmark et al. 2002), population bottlenecks associated with a pathogenic lifestyle are thought to be the main cause of the accumulation of IS elements in both *Bordetella pertussis* and *Yersinia pestis* (Parkhill, Wren et al. 2001; Parkhill, Sebaihia et al. 2003). An abundance of pseudogenes is considered to be a classic trait associated with reduced effective population size and is likely an early stage of genome reduction (Sallstrom and Andersson 2005), an extreme example is the intracellular

parasite *Mycobacterium leprae* (Cole, Eiglmeier et al. 2001). Additionally veritcal transmission, and the associated sequential bottlenecks, is held to be the cause of the relative abundance of IS elements in populations of *Wolbachia* (a reproductive parasite of insects) (Wu, Sun et al. 2004).

A notable exception to this pattern is that of *Prochlorococcus*, which is a major bacterial component of bacterioplankton, it is a free-living photoautotroph with a massive effective population size yet despite this, comparisons of three genomes sequences show evidence of genome shrinkage, AT enrichment and accelerated evolution. It has been proposed that these changes are adaptive (Marais, Calteau et al. 2008), the consequence of the presence of mutator phenotypes enabling colonisation of new ecological niches, in this case low light environments. Furthermore the massive effective population size means that *Prochlorococcus* can very effectively expunge even some of the more mildly deleterious mutations and so can more readily accommodate mutator phenotypes, even if the environmental challenge that necessitates them is as infrequent as once every few years. In particular this case highlights the necessity of detailed knowledge of the ecology, lifestyle and genetic details of a bacterial species when interpreting apparent evidence of 'genome degradation'.

1.2.5 – Comparative Bacterial Genomics

From its initial use as a technique for identifying protein coding open reading frames (ORFs) in newly sequenced genomes via comparison to closely related sequences, comparative genomics has become an essential tool to further understand a species and its relationship to its kin based on its genome sequence. With the advent of modern sequencing technologies there are now multiple genomic sequences available for many bacterial species or genera, allowing insight into the genetic makeup and gene content of any given species in greater detail than previously possible with techniques such as MLST (Enright and Spratt 1999) and MLSA. It had been thought that a well chosen 30-40 genome sequences would be sufficient to explore the entire gene content or 'metagenome' of all bacteria, however it has become evident that even within a single species in may take more than 140 genomes to fully represent the 'Pan-genome',

according to a study of 17 *Streptococcus pneumonia* genomes by Hiller (Hiller, Janto et al. 2007). Given this level of genetic diversity even between members of the same species it becomes necessary to identify all genes which are common at a given phylogenetic distance, termed the 'core' genome, as well as the genes which are present in one or more strains but absent from others, the 'accessory' genome (differences illustrated below, figure 1.2.5a).

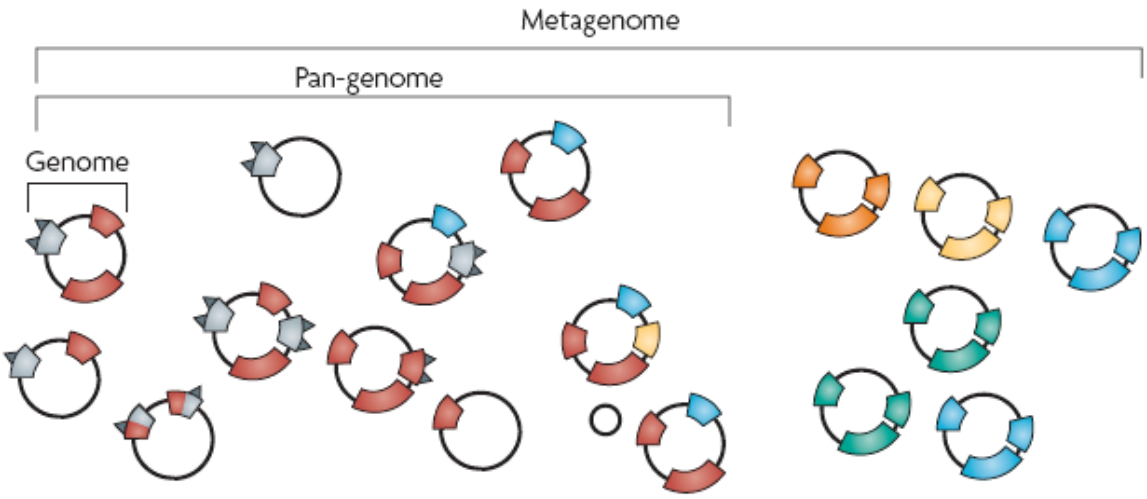


Figure 1.2.5a – A diagrammatic illustration of the differences between the Genome of a single individual, the 'Pan-genome' of a whole species and the 'Metagenome' of a microbial community. Black lines can be considered to represent 'core' genes and the coloured blocks 'accessory' genes. Adapted from Medini et al (2008).

Gain or loss of genes within an individual genome can provide insights into the evolutionary path of that strain, in that aspects of the environment or niche occupied by the strain and the challenges it faced can be inferred from its selection of genetic tools. This may include the acquisition of pathogenicity factors in the transition from a commensal to pathogenic lifestyle, or the loss or pseudogenisation of genes no longer conserved due to the establishment of a parasitic or symbiotic intracellular relationship (Andersson and Kurland 1998). Broad genome features such as inversions of sections of the genome or the patterns of base composition can also yield evolutionary information, the presence of a significantly different from average GC content or pattern of codon usage can, for example, indicate the presence of relatively recently horizontally acquired genes especially if these genes are also absent from closely related genomes (Ragan 2001).

Nucleotide differences between core genomes can provide the necessary detail to make statistically rigorous examinations of the evolutionary history between them, as mentioned above the composition of the genome in terms of bias towards G and C or A and T can be informative. In addition to this the identification of single nucleotide polymorphisms (SNPs) between orthologous sequences within the core genome provides insight into evolutionary trends which may not be evident in the extant sequence. For example two related strains/clones may exhibit highly similar genome composition in terms of %GC, with differences far less than 1% but the pattern of nucleotide changes they exhibit may show a significantly stronger bias in one genome towards a more GC poor or GC rich fate than its relative. In this way it is possible to examine how the differences in the evolutionary paths two different strains are currently on, their ‘evolutionary heading’.

1.3 – *Shigellae*

1.3.1 – Ecology of the *E. coli* as a whole

E. coli is primarily a commensal inhabitant of the vertebrate gastrointestinal tract; the initial inoculation of the GI tract comes within the first few days after birth usually either from the birth canal or by acquisition from a close family member or the environment. Whilst some strains of *E. coli* are able to survive in the environment (Lau and Ingham 2001) (Fremaux, Prigent-Combaret et al. 2008) and a few are capable of replicating (Ishii, Ksoll et al. 2006), albeit very slowly, the vast majority require a host environment for replication.

The precise composition of the gut *E. coli* population is in a constant state of flux, with strains rising to dominance and in turn being superseded by more well adapted strains. In the large intestine *E. coli* is a minority component of the biota however, it is a dominant organism in the sparsely populated lower small intestine. However it is only under disease conditions, even then only for some strains (EXAMPLE), that *E. coli* represent anything more than 1% of the bacteria identified in a faecal sample.

Pathogenic strains of *E. coli* (the *Shigellae* included) normally infect via the colonisation of mucous membranes, causing various forms of diarrhoea or urinary tract infections. In cases where epithelial disruption provides access to the blood, some strains can cause peritonitis, sepsis or Gram-negative pneumonia.

1.3.2 – Discovery and Classification of the *Shigellae*

The first of the *Shigellae* isolated was isolated in 1898 by Kiyoshi Shiga in a Tokyo hospital during an outbreak of dysentery. He isolated and cultured a bacillus which he characterised as being Gram negative, which fermented dextrose, was negative in the indole reaction and didn't form an acid from mannitol (Shiga 1898). This bacterium he named *Bacillus dysenteriae*, and has more recently been reclassified as *Shigella dysenteriae* (specifically serotype I), further studies by Shiga characterised the toxins produced by the bacterium (Shiga 1906) including the posthumously named Shiga toxin.

After Shiga's initial publication many similar organisms were identified (Flexner and Barker 1900), subsequent work identified a total of four groups of this organism and in the 1930 edition of Bergey's Manual of Determinative Bacteriology (Bergey and Bacteriologists 1930) were grouped together under the genus *Shigella*, to honour Shiga. The four species were termed; *S. dysenteriae*, *S. flexneri*, *S. boydii* and *S. sonnei*, again to honour the key workers in the field at the time – Flexner, Boyd and Sonne (Niyogi 2005).

The *Shigellae* proved to be practically indistinguishable from *Escherichia coli* by DNA – DNA hybridisation techniques; the ability to cause dysentery is a key identifying feature. However there are some strains of *E. coli* that can also cause diarrhoea and/or dysentery, specifically the Enteroinvasive *E. coli* (EIEC), which are also biochemically similar to the *Shigellae* with one of the few biochemical distinctions between *E. coli*, EIEC and *Shigella* being their metabolism of Mucate and Acetate, although this is not clear cut – 90% of typical *E. coli* are positive for both, EIEC can be positive for only one or both and *Shigella* are, with rare exceptions, negative for both (Bopp, Brenner et al. 2003).

More recently the *Shigellae* have been classified serologically into four main serogroups based upon the O antigen of the cell wall, outer membrane lipopolysaccharide (Niyogi 2005); *S. dysenteriae* (serogroup A, with 13 serotypes), *S. flexneri* (serogroup B, with 15 serotypes), *S. boydii* (serogroup C, with 18 serotypes) and *S. sonnei* (serogroup D, with 1 serotype). *S. sonnei* can also be differentiated biochemically from the other *Shigellae*, as it is positive for the beta-D-galactosidase and ornithine decarboxylase reactions.

1.3.3 – Lifestyle Choice

The *Shigellae* have adopted a somewhat different lifestyle to the other *E. coli*, with the notable exception of EIEC; in that they invade the gut lining of the human gut via Membranous epithelial cells (M Cells), whose primary function is the sampling and transport of antigens in the gut lumen (Neutra, Frey et al. 1996). The M cells take up the *Shigellae* via endocytosis and transport them across the highly impermeable gut epithelium (Sansonetti and Phalipon 1999), the intended purpose of this mechanism being to present gut lumen antigens to the immune system. However the *Shigellae* are taken up

by macrophages and induce macrophage apoptosis, causing release of the bacteria into the basal epithelial space and the release of inflammatory cytokines (Sansonetti and Phalipon 1999). Once they have access to the basal surface of the epithelium the *Shigellae* utilise invasion antigens and a type three secretion system (TTSS), encoded by the *ipa* and *mxi-spa* loci respectively, on the virulence plasmid (VP) (Menard, Dehio et al. 1996), to gain access to the epithelial cell cytosol. Once within a host cell the *Shigellae* replicate intracellularly and spread from cell to cell via manipulation of the actin cytoskeleton to form projections into neighbouring cells (Goldberg 2001) see below (figure 1.3.3a). The invasion of the epithelial cells triggers conversion of the epithelial cells into proinflammatory cells which release large quantities of TNFα and interleukin 8 (Hedges, Agace et al. 1995), causing further inflammation and disruption increasing bacterial access to the basal epithelium in a vicious cycle, resulting in the dysentery symptomatic of Shigellosis.

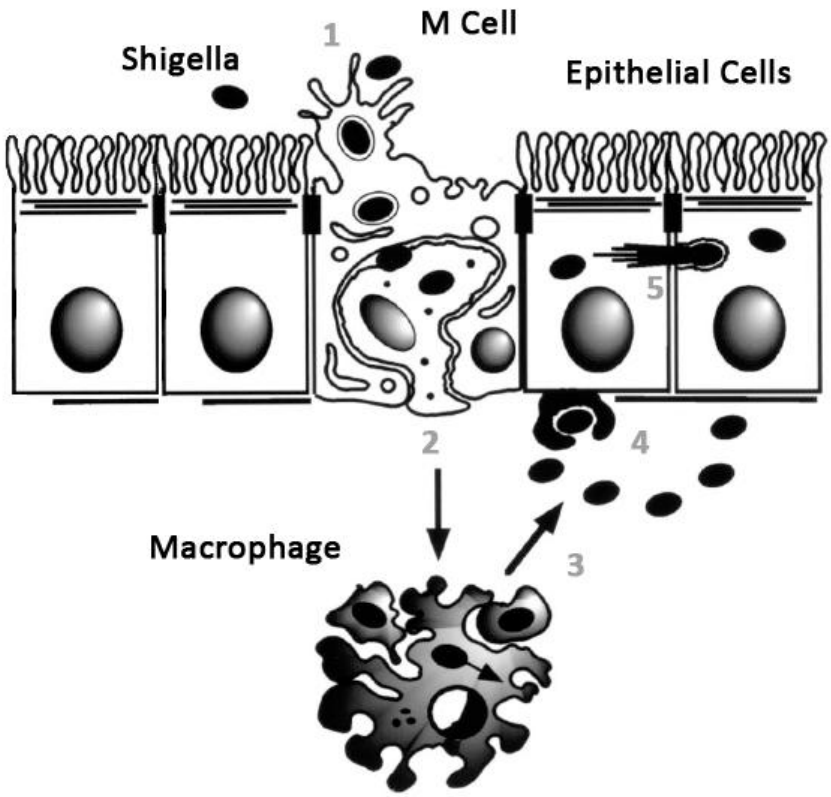


Figure 1.3.3a – Representation of the invasion of the gut epithelium by *Shigellae* adapted from (Sansonetti and Phalipon 1999). **1** – Endocytic uptake by M Cell; **2** – Transport to basolateral epithelium & Macrophage phagocytosis; **3** –Macrophage apoptotic release; **4** –TTSS invasion of epithelial cells; **5** – Intracellular replication and spread via actin manipulation.

1.3.3 – Shigellosis; Epidemiology & Diagnosis

The incubation period for Shigellosis is typically between 1 and 4 days dependant on the size of the inoculum, but can be as long as 8 days with *S. dysenteriae*. The initial symptom is watery diarrhoea; this can be the only symptom if the infection is mild. In more acute infections the disease can progress to dysentery and abdominal cramps, with the potential to result in inflammatory colitis. The disease doesn't often result in massive fluid loss (Butler, Speelman et al. 1986), however the loss of 200-300 ml of serum protein can result in severe depletion of the body's nitrogen stores, especially in cases where the infected person is already suffering from malnutrition. The Shigella are highly contagious via the faecal-oral route, with an inoculum as small as ten organisms (DuPont, Levine et al. 1989), in places where there isn't an effective sewerage system it is possible, given the low inoculum, for houseflies to act as a short-distance vector (Levine and Levine 1991).

The severity of disease caused also depends upon the serogroup or 'species' of the infecting *Shigella*, the most severe disease being caused by *S. dysenteriae* type I, as this is the only known group of *Shigellae* to produce the shiga toxin, which has enterotoxic (prevention of uptake of nutrients from gut lumen), cytotoxic (irreversible inactivation of the 60S ribosomal subunit) and neurotoxic effects (fever and abdominal cramping in primates) (Donohue-Rolfe, Acheson et al. 1991), followed by *S. flexneri* then *S. sonnei* (Keusch and Bennish 1998).

The burden of shigellosis varies from country to country and can correlate with the industrialisation and economic status of the country. Firstly the level of burden is higher in developing countries with poorer populations, around 2.1 cases per 1000 people, which rises to 13.2 per 1000 for children under five (von Seidlein, Kim et al. 2006) compared to the levels in more developed countries of around 3.2 per 100,000 people (van Pelt, de Wit et al. 2003). This can primarily be attributed to lower quality sanitation, especially in light of the ease of transmission, in more developed countries the transmission route is more readily broken with more rigorous sanitation systems.

The species or serogroup composition of *Shigellosis* cases also varies greatly between locations, and can vary over time (Dutta, Rajendran et al. 2002), with the general balance being that *S. flexneri* is the main causative agent in the developing world and *S. sonnei* is the primary agent in the developed world (Kotloff, Winickoff et al. 1999) having been observed to account for 71.7% of all cases in the USA (Gupta, Polyak et al. 2004). This has been further confirmed in a study looking at six Asian nations; Bangladesh, China, Pakistan, Indonesia and Vietnam all showed a predominance of *S. flexneri* whereas in rapidly industrialising Thailand there was a surfeit of *S. sonnei* cases (von Seidlein, Kim et al. 2006). This prevalence of *S. sonnei* in more industrialised nations is consistent with observed differences to the rest of the *Shigellae*, in that it can be transmitted via contaminated water (Leclerc, Schwartzbrod et al. 2002) or uncooked food, and has been shown to survive in an *Acanthamoeba* host (Jeong, Jang et al. 2007; Saeed, Abd et al. 2008).

The relative scarcity in isolation (von Seidlein, Kim et al. 2006) of the causative agent of the most severe form of shigellosis (*S. dysenteriae*) is consistent with its observed epidemiology as an epidemic or pandemic strain, which would preclude a low level of standing infection as any moderate level of infection could quickly expand to an epidemic or pandemic. In keeping with this it is not unsurprising to note that *S. dysenteriae* has recently been observed surviving outside a human host, in an environmental host of *Acanthamoeba castellanii* (Saeed, Abd et al. 2008), providing it with an alternative niche between epidemics.

S. boydii is rarely isolated in most parts of the world, with the exception of the Indian subcontinent as highlighted in a recent pan Asian study which reported that approximately 25% of the Shigellosis cases in Bangladesh were caused by *S. boydii* (von Seidlein, Kim et al. 2006).

Shigellae (which are non-motile, non-lactose fermenting gram negative aerobes) are identified in the laboratory by first aerobically culturing faecal samples on plates of MacConkey agar, the aerobic conditions inhibit the growth of the gut microbes which are

obligate anaerobes, bile salts in the medium inhibit the growth of Gram positive bacteria in the sample and neutral red dye highlights colonies fermenting lactose, after 24 hours slants of Triple Sugar Iron agar are stabbed with the colonies identified as non lactose fermenting. *Shigellae* (and a small number of *E. coli*) will show an alkaline slant and acidic butt (deep in the stab) after 18 hours of incubation. This is a result of their metabolism being restricted to glucose, which under aerobic conditions is exhausted and so the bacteria switch to the consumption of peptones producing ammonia. In the butt of the stab, the pyruvate from anaerobic metabolism of the glucose is converted to acidic end products, without the production of any gas.

This is sufficient for a presumptive diagnosis of a *Shigella* infection (Hormaeche and Peluffo 1959); however there are a few *E. coli* strains which also display these results; differentiation is achieved by testing for the ability to decarboxylate lysine – the *Shigellae* have a non-function Lysine Decarboxylase. Full identification of the exact strain or serotype can be achieved via agglutination assays with *Shigella* antisera.

1.4 – Aims of the Project

Whilst the genomes of many closely related bacterial species have been sequenced there have been few studies on the characterisation of short term evolutionary trends, especially in cases where a recent adoption of a differential ecological niche has yet to have a pronounced effect on the extant composition of the genome.

The aim of my project was to explore the time dependence of purifying selection previously observed (Rocha, Maynard Smith et al. 2006), using the patterns of nucleotide changes observed between genome sequences of members of the same 'species' and relate and patterns or trends observed to the population structure of the individual species/strains analysed.

To that end the main focus is on an examination of several clones of *Escherichia coli*, 5 nominative *E. coli* and 4 *Shigellae* including a representative strain of each of the four named *Shigella* species – *boydii*, *dysenteriae*, *flexneri* and *sonnei*. The *Shigellae* have all adopted a facultative intracellular pathogenic lifestyle, whereas the pathogenic *E. coli* included in the analysis do not reproduce inside host cells. Analysis of the nucleotide and amino acid polymorphisms present in the 'core' genome sequences revealed the time-lag effect associated with purifying selection and highlighted a notable and statistically significant difference between *E. coli* and *Shigellae*, congruent with their adoption of an intracellular lifestyle. This trend was further examined to identify in more detail the key factors responsible for the trend.

Chapter 2 – Materials & Methods

2.1 – Genomic data used in this project

2.1.1 – The strains and species included in the datasets

The primary dataset comprises nine genomes from the *Escherichia coli* group, five of which are formal *E. coli* and include two Uropathogenic strains (CFT073 & UTI89), one Enteroaggregative strain (EAEC042), one Enterohaemorrhagic strain (O157:H7 Sakai) and one laboratory reference strain (K-12 MG1655). The other four sequences are representatives of the four nominal *Shigella* species; *boydii* (Sb277), *dysenteriae* (Sd197), *flexneri* (2a301) and *sonnei* (Ss046). These genomes were all chosen as they represent a large portion of the pathogenic *E. coli* and also include four genomes which have relatively recently adopted a different ecological niche (facultative intracellular pathogenesis in the *Shigellae*).

Code	Full		Accession Number	Reference
	Species	Strain		
EcA	*Escherichia coli*	K-12 MG1655	U00096	(Blattner, Plunkett et al. 1997)
EcB	*Escherichia coli*	O157:H7 Sakai	BA000007	(Makino, Yokoyama et al. 1999)
EcC	*Escherichia coli*	CFT073	AE014075	(Welch, Burland et al. 2002)
EcD	*Escherichia coli*	EAEC042	*Sanger Centre*	*www.sanger.ac.uk*
EcE	*Escherichia coli*	UTI89	CP000243	(Chen, Hung et al. 2006)
Sb	*Shigella boydii*	Sb227	CP000036	(Yang, Yang et al. 2005)
Sd	*Shigella dysenteriae*	Sd197	CP000034	(Yang, Yang et al. 2005)
Sf	*Shigella flexneri*	2a301	AE005674	(Jin, Yuan et al. 2002)
Ss	*Shigella sonnei*	Ss046	CP000038	(Yang, Yang et al. 2005)

Table 2.1.1a – A list of the Species and strains used along with the short codes used and the accession numbers of the genome sequences and where relevant, references.

2.1.2 – Identification & alignment of all orthologous genes

Genomes were compared in a pair-wise fashion. For a given pair of genomes, say A and B, the top ten BLAST hits for each gene from genome A in genome B and vice versa were identified and 'shortlisted', i.e. reciprocal best hits. These hits were then refined via an exact pair wise alignment with the Needleman-Wunsch dynamic programming algorithm, with end gaps not counted negatively. Genes were considered orthologous if they were reciprocal best hits in the dynamic programming alignment with greater than 40% amino acid sequence similarity and less than 20% difference in length. Finally genes lying outside conserved syntenic blocks were considered possible false orthologues and were discounted.

The core genome for a given set of sequences was considered to be the intersection of all the pair-wise identified orthologous genes. These 'core' genes were translated and aligned by amino acid sequence, using clustal W (Thompson, Higgins et al. 1994), before being back-translated to nucleotide sequences, this ensured maximal accuracy of the alignment.

The genes were output to individual FASTA files, which were then used to produce a single sequence for each taxon; first the files were concatenated to produce a single FASTA file with individual entries for each gene. Using '*groupcat'*, written by Eduardo Rocha of the Institut Pasteur in Paris, these sequences were then concatenated into a single alignment, with each taxon represented by one long sequence. However, this alignment does not precisely share the same syntenic structure as the full genomes due to gene order variation between the strains.

2.2 – Phylogenies

2.2.1 – Neighbour-Joining

Neighbour-joining trees were generated using the 'MEGA' (Molecular Evolutionary Genetic Analysis) software package, version 4.0 (Tamura, Dudley et al. 2007), with the Kimura 2-parameter model, assuming no among-site rate variation and no pattern variation among lineages and a thousand bootstrap replicates, based upon the concatenated alignment sequences described above (2.1.1).

2.2.2 – Bayesian Analysis

MrBayes uses Bayesian approaches to calculating probabilities within a metropolis-coupled markov chain monte carlo algorithm to estimate the posterior probability of trees given the input multiple sequence alignment, as a means to explore tree-space to locate the optimal (or a most-optimal) tree, whilst simultaneously estimating the parameters of the evolutionary model specified as the mechanism of sequence change, using each new 'most optimal' tree as the basis for the next step, and so incrementally improving the tree. The program implements this via the use of several 'chains' exploring tree-space in parallel, usually in the ratio of 3 'heated' chains to 1 'cold' chain. The heated chains acting as 'scouts' scanning tree space and the cold chain making a more thorough search, the heated and cold chains are interchanged at specified intervals based upon likelihood values, this allows the escape from local optima in the search for the global optimum (or optima).

The neighbour-joining trees were verified by Bayesian analysis, however given the high computational cost of the Bayesian method, analysis of entire concatenated alignments was unfeasible, thus 50 20Kbp segments of the alignment were used in order to provide a sufficient sampling of the sequence data to be reliable yet in sections small enough to be analysed in a realistic timeframe. Both a random selection and an ordered selection of these segments were made. The figure below (Figure 2.2.2a) shows the distribution of these segments along the core *E. coli* / *Shigella* genome, only the first 1Mbp of the alignment is shown, for illustrative purposes.

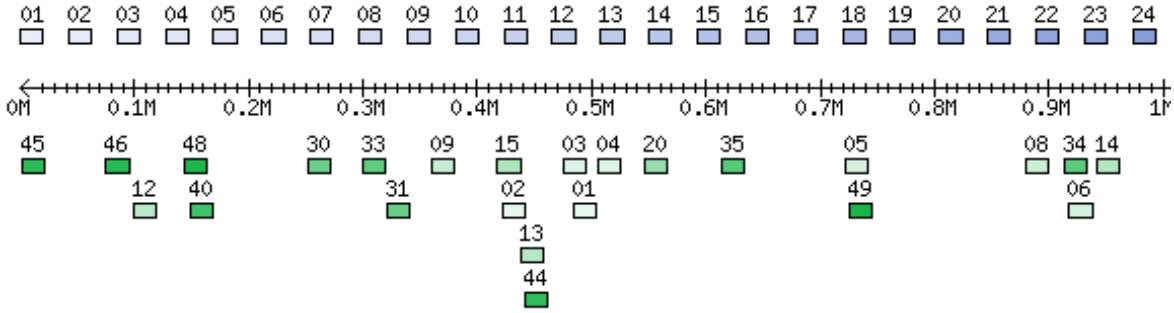


Figure 2.2.2a – The distribution of the segments selected for Bayesian analysis, along the first 1Mbp of the *E. coli* / *Shigellae* alignment, the random segments in shades of green, the ordered in shades of blue.

These segments were each run independently using the configuration in the Bayes block (shown below, first code paragraph). A standard nucleotide substitution model was specified, operating under the general time reversible model with a gamma distribution of rate variation among sites and a proportion of invariable sites. Also specified are the parameters of the analysis as well as commands to enable automatic conclusion of the analysis on the basis of the 'average standard deviation of split frequencies', this enabled the entire process to be automated via a BASH script (shown below, second code paragraph).

```
begin mrbayes;
     lset nucmodel=4by4 nst=6 rates=invgamma;
     mcmcp samplefreq=10 ngen=100000 relburnin=yes
     burninfrac=0.25 stoprule=yes stopval=0.001;
     mcmc;
     sump burnin=250;
     sumt burnin=250;
end;
```

[Bayes Block used in the analysis]

```
#! /bin/bash

for file in *.bayes.nex;
do

     while [ "`ps -U Kev | grep mrbayes | wc -l`" -gt 4 ]

     do
      sleep 15
     done

     time=`date +%H:%M`

     echo "$file Started -> $time";
     screen -dmS $file mrbayes $file;

done:
```

[BASH Script, automating the MrBayes analysis ensuring that no more than 4 analyses were running in parallel for the user 'Kev', and starting each new process in an independent shell]

All 50 trees were used to produce a 50% majority rule consensus tree, treating the input trees as unrooted, as implemented in the program *'consense'* part of the PHYLIP package (Felsenstein 1989).

2.3 – Base Counts

The absolute number of all four nucleotide bases in the concatenated sequences of the aligned orthologues was determined via a perl script written by the author, which returned overall base counts, as well as the counts at fourfold degenerate sites, non-fourfold degenerate sites and all three codon positions independently. Gaps and ambiguously coded bases (R, S, Y, M etc) were reported by the script but not included in the final counts.

2.4 – SNP Analysis

2.4.1 – SNP Selection

In order to be conservative only those SNPs which corresponded to a conserved (i.e. monomorphic) site in all the other genomes were included (i.e. unique singleton SNPs). Polarisation of the polymorphisms (e.g. A > B or B > A), essential for normalisation as well as enabling the separation of AT enriching SNPs from GC enriching SNPs, was achieved parsimoniously by comparison of the unique singleton polymorphism to the conserved base in the other genomes, illustrated below (figure 2.4.1a). This approach has been shown to be consistent with maximum likelihood estimations (as in PAML) over very short divergence times (Hurst, Feil et al. 2006). The short divergence times involved in the analysis and relative scarcity of SNPs in the majority of the sequences (one per 0.5-12 Kbp) means that multiple hit errors are largely insignificant. This was confirmed by comparison of the number of SNPs observed to the expected value, determined using the probabilistic approach outlined below (2.11.3). Counts were performed, as with base counts, at all sites as well as, fourfold degenerate sites, non-fourfold degenerate sites and 1^{st} 2^{nd} and 3^{rd} codon positions independently, using the program '*subst*' written by Eduardo Rocha of the Institut Pasteur in Paris.

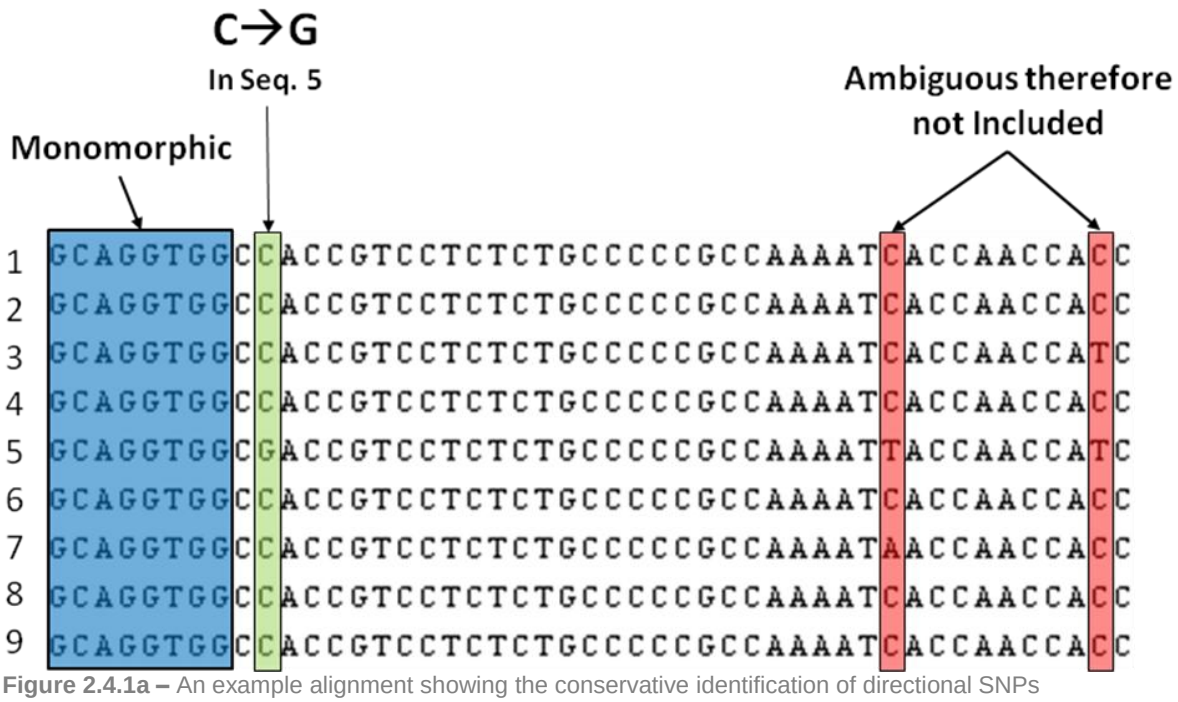


Figure 2.4.1a – An example alignment showing the conservative identification of directional SNPs (highlighted green) along with sites with SNPs where direction ambiguous (highlighted in red) and example monomorphic sites highlighted in blue.

2.4.2 – Normalisation of SNP Counts

In order to directly compare any two sets of SNP counts, allowances need to be made for any differences between them that may simply be a reflection of their different origins. Here it is mainly the nucleotide composition of the genome that varies, inequality in the base composition means that SNPs originating in any given base are only comparable to other SNPs originating from the same base. That is, there is an increased likelihood of an A→X change over a G→X change if the genome contains a surfeit of A's over G's. Base composition, however, does not account for the differences in the frequencies of the SNPs originating from the same base, for example the difference between T→A and T→C.

To correct for this each SNP type is divided by the count of its originating base, e.g. A→T divided by the number of A's. This gives the number of any given change type per originating nucleotide in the genome and makes all the SNP types directly comparable.

These values aren't comparable between taxa, owing to the different number of SNPs observed in each genome. To correct for this, each set of 'SNP counts per originating base' is scaled such that they sum to one, by dividing the value for each SNP type by the sum of all 12 SNP types for that genome. This yields a set of SNP values whereby each SNP is expressed as a proportion of all the SNPs observed, allowing for unequal base

composition within each genome. This set of values is referred to as a *'Polymorphism Profile'*.

2.4.3 – Estimation of Time Associated with Polymorphism Profiles

To compare the polymorphism profiles of different genomes and allow for different ages of profile, it is necessary to estimate the age of the polymorphism profile. This was achieved by assuming that the SNPs within a genome were evenly distributed along the branch of that genome within the phylogenetic tree (i.e. over time), and then the median 'age' of the SNPs, in 'number of SNPs before the present', was taken to be the midpoint of the branch. The midpoint is simply calculated as half of the total number of SNPs observed in the genome, this was then log transformed to linearise the data.

2.4.4 – Calculation of Metric Ratios

Two readily calculated ratios were used to summarise the polymorphism profiles, allowing for easy visualisation of trends within sets of closely related genomes. The ratio of AT enriching to GC enriching SNPs and the ration of Transitions to Transversions, were chosen as they are two independent methods of subdividing nucleotide changes. The ratios were calculated as below, using the polymorphism profiles associated with all sites, fourfold and non-fourfold degenerate sites and 1st 2nd and 3rd codon positions independently.

$$+AT/+GC = \frac{CA + CT + GA + GT}{AC + AG + TC + TG} \qquad Ti/Tv = \frac{AG + CT + GA + TC}{AC + AT + CA + CG + GC + GT + TA + TG}$$

Equation 2.4.4a – The calculation of +AT/+GC and Ti/Tv metric ratios from the normalised SNP data in Polymorphism Profiles, where CA refers to the normalised value for C to A changes

2.4.5 – Determining confidence intervals for observed ratios

Given that any set of SNP counts is simply a small sampling of the overall SNP bias associated with the genome under analysis, it becomes necessary to estimate how confident it is possible to be in the pattern of SNP counts observed and any metrics derived from them.

To that end, the absolute SNP counts for a given genome were resampled, with replacement, multiple times (typically 1000) producing replicate sets of SNP counts. These bootstrapped SNP counts were normalised as above (2.4.2), using the original base counts of the genome in question, various metric ratios were then calculated from the resultant polymorphism profiles. The values from all of the bootstrap replicates were then used to estimate the confidence interval for those ratios for the given genome.

2.4.6 – Bootstrap Analysis of Metric Ratio Differences

Comparison of any two polymorphism profiles is hampered by the fact that they represent different sample sizes of the underlying SNP biases from their respective genomes. In order to accurately compare the two, it becomes necessary to resample one of the polymorphism profiles such that both are of the same size.

The absolute SNP counts for two genomes, one as the 'reference' genome (A) and one as the 'comparator' genome (B) were used and the comparator (B) resampled, with replacement, multiple times (typically 1000-10000) such that the total number of SNPs in the bootstrapped B SNP counts (B') was equal to that in the A SNP counts. Each of the B' sets of SNP counts was normalised as above (2.4.2) and metric ratios calculated from the resultant polymorphism profile. The ratios from A were the calculated and converted to percentiles of the B' values, with percentile values greater than 95% and lower than 5% representing a 95% confidence that A is higher or lower (respectively) than B

2.4.7 – Determination of within-branch dN/dS

The polymorphisms in a given genome were identified as above and their location within the sequence noted, this was repeated for all genomes in the dataset. These lists of polymorphism locations, types and directions were then used to 'revert' each genome to an inferred ancestral state. MEGA version 4 (Tamura, Dudley et al. 2007) was then used, to determine the dN/dS ratio, using the Nei-Gojobori model (Nei and Gojobori 1986) with a Jukes-Cantor correction, between the original sequence for each genome and it's inferred ancestral sequence, in essence yielding the dN/dS ratio of the SNPs in the polymorphism profile.

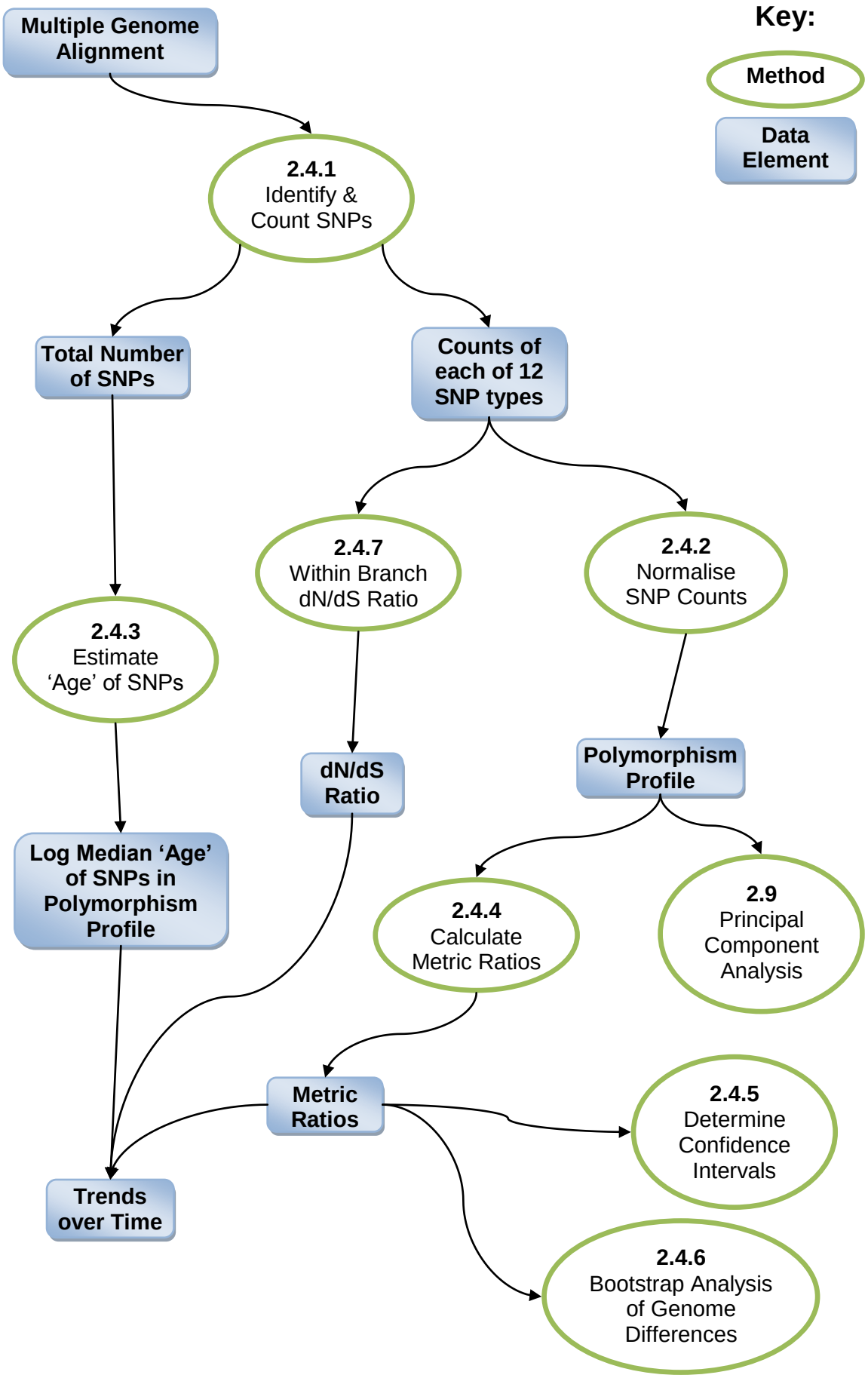


Figure 2.4a – A Graphical Summary of the Nucleotide analysis process, with methods in the green ellipses and elements of data in the shaded grey boxes

2.5 – Amino-Acid Polymorphism Analysis

2.5.1 – Identification

Amino-acid polymorphisms were identified using the same conservative criteria as those used for the identification of nucleotide polymorphisms, again polarisation of the changes was determined on a parsimonious basis, from 'majority conserved' amino acid to 'singleton' amino acid, using the program '*dna2aasubst*' written by Eduardo Rocha of the Institut Pasteur in Paris. The total numbers of gains and losses for each amino acid were summarised and output as the values of Gains plus Losses and Gains minus Losses for each amino acid.

2.5.2 – Normalisation

In order to compare the rates of change in any given amino-acid the number of changes had to be normalised to make them comparable; for each amino-acid the net number of changes (Gains minus Losses) was calculated it was then divided by the total number of changes involving that amino acid (Gains plus Losses) to yield the net change per amino-acid for each amino-acid.

2.5.3 Metabolic Costing

The net change for any given amino acid in a given genome, e.g. Y in Genome 1, was multiplied by the metabolic cost of that amino acid (see section 4.1.2), to yield a net metabolic cost of all changes for Y. This was calculated for each amino acid, and the sum of all costs divided by total number of unique amino acid polymorphisms (i.e. the sum of all the Gains+Losses counts divided by two). This gave the mean metabolic cost per amino acid polymorphism in each genome, in high energy phosphate bond equivalents.

2.6 – Taxon Exclusion Analysis

Additional branch lengths for a given genome were inferred by systematic exclusion of the other taxa either individually or in groups, always keeping a minimum of three taxa. This 'pruned' dataset was then reanalysed. These reanalysed points are then used as inferred additional timepoints for each taxon. This method was initially used to ensure accuracy in

the regression analysis (below) however given the non-independence of the multiple points for each genome the method was superseded by the use of the internal node/branch analyses. This method is only used during the Genomic AT content simulations where multiple timepoints for a single genome are necessary.

2.7 – Trend over time Plot & Residual Analysis

The direct comparison of any given metric between genomes was corrected for the effects of the differing branch length (i.e. time), by plotting the values of the metric against the Log divergence time. The residual, for each point, to the overall regression line was then taken to represent the metric for the corresponding genome. Comparison of these values was then used to determine any time-independent effects evident in the metric used.

2.8 – Internal Branch Analysis

2.8.1 – Calculated (Terminal Branch Subtraction – TBS)

The polymorphism profiles associated with internal branches was estimated via terminal branch subtraction. For example for a tree of topology (D(C(B,A))) the absolute numbers of SNPs on the terminal branches leading to B and A are counted as above, the pattern of SNPs associated with the internal branch leading to the node supporting both B and A is estimated by recounting the SNPs associated with B when A is excluded (B') and subtracting the SNPs counted for the terminal branch leading to B. This is then repeated for A' minus A and the mean of the two values taken. For deeper branches a mean was taken of the values for all successive branch exclusion and subtractions. These SNP counts are then normalised as above, using the mean base counts of all the terminal branches supported by the internal branch, to yield the polymorphism profile associated with a given internal branch. A worked example is shown below in figure 2.8.1a.

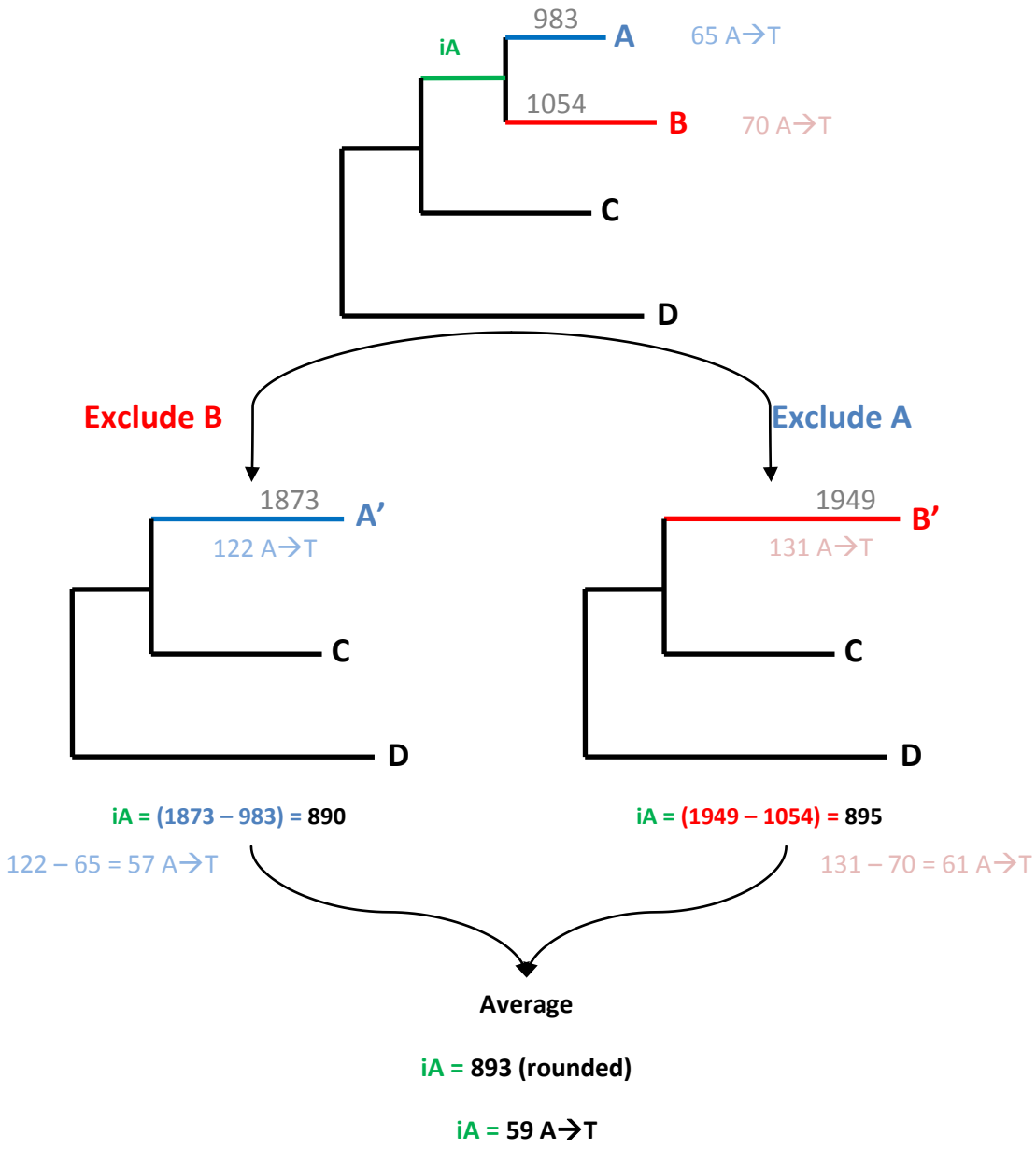


Figure 2.8.1a – A worked example of the TBS approach to internal branch SNP calculation, showing both potential routes to the same internal branch, for deeper trees the principle is the same, however the application is more complicated. Pale colours show a worked example for a specific directional SNP type.

2.8.2 – PAML

Using the *codeml* part of the PAML application suite, the sequences associated with the internal nodes were reconstructed using a maximum likelihood approach on the basis of the neighbour-joining tree for the data set. The exact model implemented in *codeml* is an Empirical Bayes model, which takes account of branch lengths and relative substitution rates, which are themselves estimated via a maximum likelihood approach (Yang 2007).

The analysis was run using a configuration as listed below;

```
      seqfile = data.phy
     treefile = data.tree
      outfile = data.mlc

        noisy = 9  *Screen Output Detail (0-9)
      verbose = 1  *Amount of Detail in Screen output (0-2)
      runmode = 0  *User Specified Tree

      seqtype = 1  *Codons
    CodonFreq = 2  *F3x4 Model - Use nucleotide frequencies,
                                 But consider codon positions seperately

*        ndata = 10
        clock = 0  *No clock
       aaDist = 0  *Equal distances assumed

        model = 2  *2 or more dN/dS ratios for branches

      NSsites = 0  *One ratio - largely irrelevant with ancestral
                        reconstruction

        icode = 0  *Unversal Genetic Code
        Mgene = 0  *Equal codon substitution rates

    fix_kappa = 0  *To be estimated
        kappa = 2  *Initial value
    fix_omega = 0  *To be estimated
        omega = 0.4 *Initial Value

    fix_alpha = 1  *Fixed Alpha
        alpha = 0. *Infinity / Constant Rate
       Malpha = 0  *Same alpha for all genes

        getSE = 0  *Don't retrieve standard error of Estimates
 RateAncestor = 1  *Reconstruct Ancestral Sequences

   Small_Diff = .5e-6
    cleandata = 1  *Remove Sites with ambiguity data
*  fix_blength = -1  *Use random starting points for ML estimation of
                        Branch lengths
       method = 0  *Simultaneous Optimisation of all parameters
```

[Codeml.ctl File contents specifying the configuration of the ancestral sequence reconstruction]

These ancestral sequences were then processed as above (2.4.1 – 2.4.4) excluding the terminal branches which on the tree are supported by the internal node associated with the inferred ancestral sequence.

2.8.3 – Estimating Time associate with Internal Branch Polymorphism Profile

Based upon the earlier method (2.4.3) the divergence time associated with the polymorphism profile for an inferred 'ancestral' internal node sequence was estimated as follows:

The mean distance from any given SNP in the internal branch "iA" shown below to the tips of the tree (extant sequences), is simply the distance from that SNP to the node supporting taxa A and B, plus the mean distance from that node to the present, which here is the mean of the branch lengths leading to A and B (figure 2.8.3a). Therefore the average distance from any given SNP in the internal branch iA to 'now' can be simply estimated; assuming the SNPs are evenly distributed along iA, the average (median) distance to the internal node supporting A and B is just the total number of SNPs in iA divided by two, plus the distance to the end of the tree.

This was then log transformed to linearise the data, in line with the approach in 2.4.3, thus the divergence time was estimated as the log of; half the SNPs in the internal branch plus the mean of all supported branches.

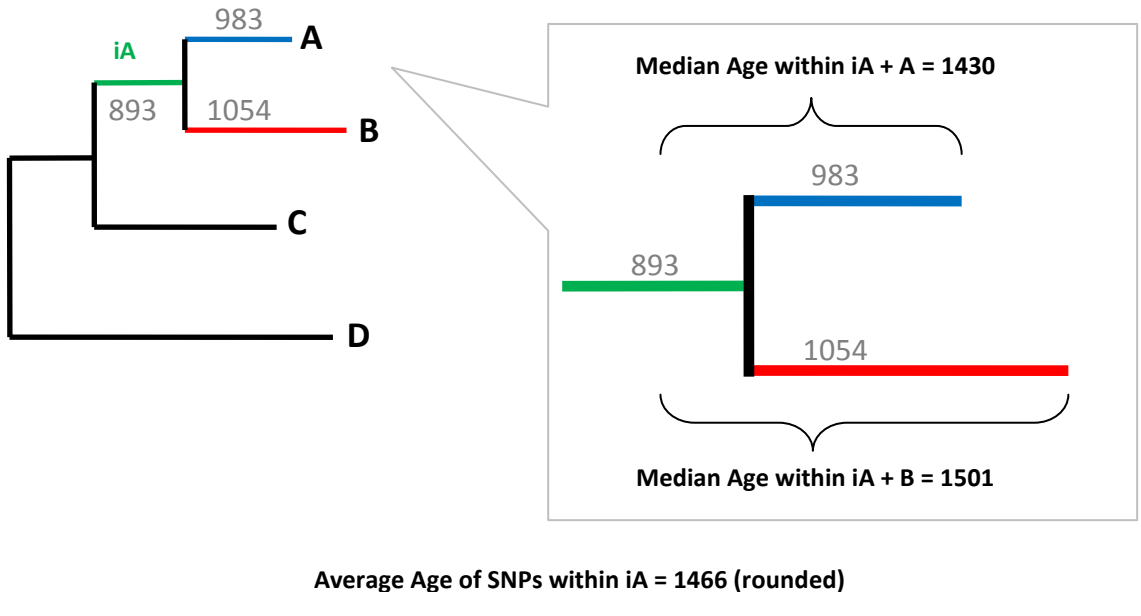


Figure 2.8.3a – A sample phylogeny, with relevant internal and terminal branches highlighted to illustrate the method of estimating the divergence time associated with an internal branch polymorphism profile

2.9 – Principal Component Analysis

Identification of trends and patterns present in data is troublesome, prone to bias and can be especially difficult for multidimensional data where visualisation is problematic. In order to analyse multidimensional datasets it is necessary to reduce the dimensionality of the data, whilst retaining the majority of the information. Principal component analysis re-explains multidimensional datasets using multiple uncorrelated variables (principal components); each successive variable explains the majority of the remaining unexplained variation. The analysis also permits the breakdown of the contribution of the initial variables to each principal component, allowing the variation captured by any principal component to be explained as ratios of the original input variables.

Rigorous statistical comparison of the polymorphism profiles was achieved using principal component analysis, providing 12 input variables (the 12 SNP types) and recovering all 12 possible principal components using Minitab version 15 (Minitab 2006). For each of the 12 principal components the relative weightings for each of the input variables was calculated as part of the analysis. Through examining these weightings it is possible to determine what real world factor(s) have the most influence within each principal component. Any given principal component can also be described by the ratio of its most positively and most negatively weighted input variable.

2.10 – Sliding Window Analysis

2.10.1 – Base Counts

Base counts were performed on the gene sequences prior to concatenation, summing the base counts over groups of genes (typically 100), with the group to be summed sliding along the genes by typically $1/10^{th}$ of the group length. The total number of bases at each codon position as well as fourfold, non-fourfold and all sites was counted, to enable correct normalisation of the corresponding SNP counts.

2.10.2 – SNP Counts

In a similar method to above (2.10.1) SNP counts were summed using sliding groups, of the same size and with the same group step as the corresponding base counts. Once again the SNP counts were calculated for those at each codon position as well as fourfold, non-fourfold and all sites. These SNP counts were then normalised as above (2.4.2) to yield polymorphism profiles associated with each group of genes.

2.10.3 – SNP Density

Variations in SNP density were calculated from the concatenated aligned gene sequences of the original dataset such that each species or strain was represented by a single large sequence. SNPs were identified and counted as above (2.4.1), within sliding windows, typically 10Kbp, with these windows sliding typically $1/10^{th}$ their size each time.

The raw SNP counts were normalised to give the difference between the mean number of SNPs per window and the number of SNPs observed in a given window, expressed as proportions of the total number of SNPs in the sequence for a given species or strain.

2.11 – Simulated Equilibrium Genomic AT Content

Whilst the pattern of SNPs unique to a taxon can provide insights into the current mutation or selective biases, in concert with other information it can be used to examine the likely mutational or selective equilibrium of the taxa. Specifically the pattern of nucleotide changes can be used to estimate the equilibrium base content of the genome. The methods below provide three different approaches to this task. The first two models approximate the mutational equilibrium of the genome base composition using the observed pattern of nucleotide changes; the third approximates the selective equilibrium using a dynamic pattern of nucleotide changes inferred through use of the taxon exclusion analysis (See 2.6).

The assumptions behind the simulation are graphically described in figure 2.11a. The presence of a mutation or selection bias produces a bias in the pattern of SNPs, of which the observed SNPs are a sample. Through the simulations the data in the observed SNP patterns can be combined with the extant genome composition to estimate the equilibrium

value for the genome, given the pattern of SNPs. The estimated equilibrium is then an approximation of the original mutation or selection bias.

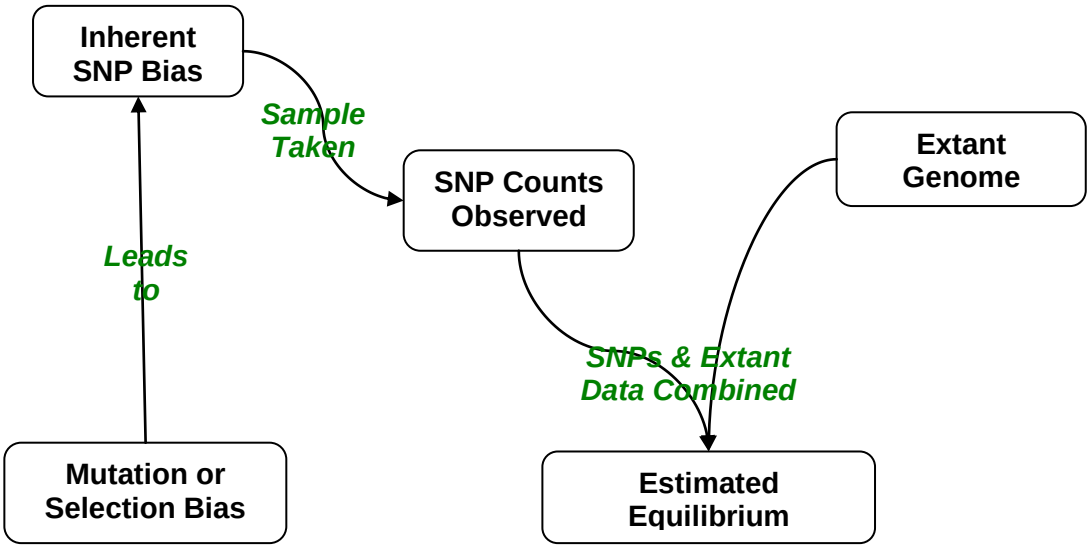


Figure 2.11a – A graphical summary of the reasoning and processes behind the Genomic AT Simulation techniques.

2.11.1 – Simple Stepwise Model

Probabilities were defined from the base counts for the genome being simulated and used to select, at random, a base e.g. 'A'. This base was then assumed the be the location of a new mutation event and using 'A' as the originating base a SNP type was selected at random, using probabilities defined from the raw SNP counts of the SNPs of the type A→X. Once selected this SNP type was applied, taking the example of A→T, the count of base A was reduced by one and the count of T increased by one. This was then iterated until the proportion of A+T reached a plateau (typically several million iterations).

2.11.2 – Matrix of Substitutions

The number of absolutely conserved sites for each base, along with the raw counts of each of the SNP types was used to create a matrix representing the rates of base interchanges in the genome being analysed, whereby the matrix is arranged such that each column represents an originating base, and each row a destination base, in alphabetical order. Thus column 1(A), row 3(G), would contain the value representing the proportion of A's which were observed to change to G's. This is used as a

transformational matrix and applied to a column vector containing the base counts for the genome.

Each iteration of the simulation is a single vector-matrix multiplication, a worked example of this is shown below:

$$\begin{bmatrix} A \\ C \\ G \\ T \end{bmatrix} \times \begin{bmatrix} AA & CA & GA & TA \\ AC & CC & GC & TC \\ AG & CG & GG & TG \\ AT & CT & GT & TT \end{bmatrix} = \begin{bmatrix} (A \times AA) + (C \times CA) + (G \times GA) + (T \times TA) \\ (A \times AC) + (C \times CC) + (G \times GC) + (T \times TC) \\ (A \times AG) + (C \times CG) + (G \times GG) + (T \times TG) \\ (A \times AT) + (C \times CT) + (G \times GT) + (T \times TT) \end{bmatrix} = \begin{bmatrix} A' \\ C' \\ G' \\ T' \end{bmatrix}$$

Equation 2.11.2a – A C G & T represent the absolute counts of the appropriate bases, XY represents the proportion of base X observed to have changed to base Y.

Thus A', for example, is the sum of the unchanged As (A x AA) plus the number of As originating from mutations from other bases (N x NA). This approach allows for base loss as well as gain, by the inclusion of the conserved (NN) changes, all the changes originating from a given base will sum to one, such that the calculation of A x AA takes into account those A's lost via mutations to C G and T.

2.11.3 – Dynamic Matrix of Substitutions

In order to allow for change in the matrix over time it was necessary to have multiple timepoints for the genome being analysed. These were estimated using the taxon exclusion approach (2.6). For each matrix element (AA, AC and so on), the trend over time was estimated as a linear regression with respect to Log_{10} of the total number of changes. Using these trends and taking the 'earliest' timepoints as starting values, the matrix was then re-calculated after every iteration, using the Log_{10} of total number of changes the simulation had accrued to that point, correcting for multiple hits, i.e. the total number of observable changes accrued.

The number of new SNPs in any given iteration was calculated by performing the matrix multiplication, omitting the results of A x AA, C x CC, G x GG and T x TT, this was then added to the previous total number of SNPs to yield the total number of SNPs after the iteration.

In order to correct for multiple hits, a computationally efficient probabilistic approach was chosen, the derivation of the equation is shown below;

①	$\frac{1}{B}$	Probability of a SNP occurring at any base at random in a genome of B base pairs
②	$\left(1-\frac{1}{B}\right)$	Probability of a SNP missing any base at random
③	$\left(1-\frac{1}{B}\right)^{S}$	Probability of a base being missed, given *S* SNPs have occurred
④	$1-\left(1-\frac{1}{B}\right)^{S}$	Probability of a base sustaining a SNP, given *S* SNPs have occurred
⑤	$B \times \left[1-\left(1-\frac{1}{B}\right)^{S}\right]$	The likely number of bases where SNPs have been sustained (i.e. Observable SNPs), given *S* SNPs have occurred

Equation 2.11.3a – The derivation and explanation of the equation used to correct for multiple hits, where *S* and *B* are the number of SNPs that have occurred and the number of bases in the sequence, respectively.

The formula uses a probabilistic approach to estimate the expected number of observable SNPs (S') from S SNPs applied to a sequence of size B (bp). It is worth noting that the 'true' number of observable SNPs can only be accurately described by a distribution, however generating this distribution from repeated rounds of random SNP placement is computationally intensive. The above formula provides an estimate of the mean of this distribution and is several orders of magnitude faster. The value yielded by the equation very closely approximates the mean of the distribution of number of observable SNPs seen when simulating the random placement of SNPs in genomes of various sizes.

Chapter 3 – Patterns of Nucleotide Change in *Shigellae*

3.1 – Introduction

3.1.1 – Phylogenetic and Genomic Characterisation

Comparisons of *E. coli* and *Shigellae* by various methods such as MLEE (Ochman, Whittam et al. 1983), ribotyping (Rolland, Lambert-Zechovsky et al. 1998) and MLEE combined with single gene phylogeny of *mdh* (Pupo, Karaolis et al. 1997), have shown the *Shigellae* to fall well within the *E. coli*. This observation has since been confirmed by an eight housekeeping gene phylogenetic analysis (Pupo, Lan et al. 2000) and a more robust 23 housekeeping gene phylogeny (Yang, Nie et al. 2007). A 169 gene phylogeny of the available fully sequenced *E. coli* and *Shigella* strains also shows the *Shigellae* and *E. coli* to be highly closely related (van Passel, Marri et al. 2008).

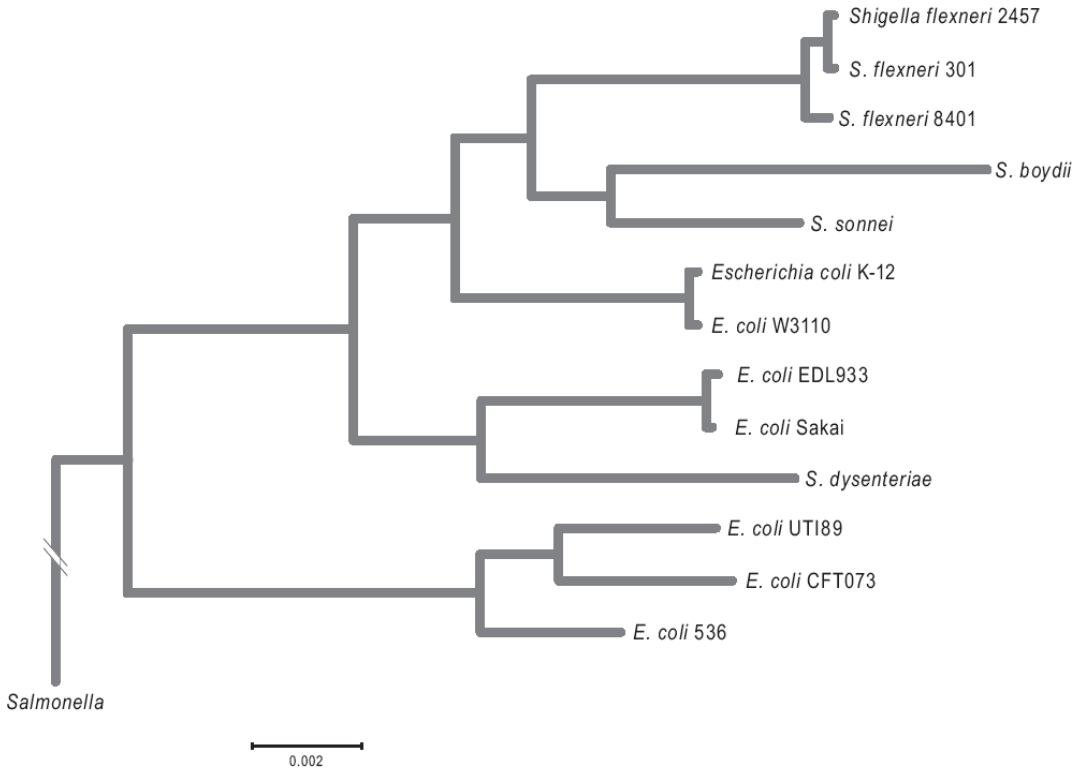


Figure 3.1.1a – Phylogenetic tree showing the *Shigellae* and *E. coli* as one interspersed clade, with *Salmonella* as an outgroup. Adapted from van Passel et al (2008)

This phylogenetic positioning is congruent with the accepted notion of the *Shigella* lineages arising through multiple independent acquisitions of the virulence plasmid pINV, somewhere between 35,000 and 270,000 years ago based upon molecular clock

estimates and depending on the lineage (Pupo, Lan et al. 2000), or as recently as 10,000 years ago for *S. sonnei* (Shepherd, Wang et al. 2000).

The *Shigellae* fall into three main clusters, phylogenetically, which bear only minimal resemblance to the nominal 'species', *S. flexneri* showing the most congruence, *S. boydii* and *S. dysenteriae* show little cohesion within the phylogenetic tree and are spread across multiple mixed 'species' clades, notably *S. sonnei* falls outside the three main clades observed (Yang, Nie et al. 2007).

EIEC strains have been shown to posess the pINV plasmid (Sansonetti, d'Hauteville et al. 1982; Silva, Toledo et al. 1982), showing multiple distinct acquisitions of the plasmid, forming a set of distinct phylogenetic clades which do not overlap with the *Shigellae*. The EIEC also show lower levels of sequence variation (Pupo, Lan et al. 2000) and lower number of O antigen variants (Lan, Alles et al. 2004), suggesting that they are a more recent set of acquisitions of pINV possibly representing an intermediate form between commensal *E. coli* and full-blown *Shigellae*.

Whole genome sequences have permitted more detailed comparisons of the *Shigellae* with the *E. coli*, revealing several consistent differences. The *Shigellae* possess a markedly greater number of insertion sequences (IS) in their genomes (Moran and Plague 2004) representing between 7 and 12 percent of their genome, compared to approximately 1 percent of the *E. coli* MG1655 genome, potentially a contributory factor to the great number of genomic rearrangements, insertions and deletions observed in the *Shigellae*, specifically *S. dysenteriae* which possesses the highest number of IS – 623. There are also a greater number of pseudogenes in the *Shigellae* (Yang, Yang et al. 2005); corresponding to 4 to 8 percent of the open reading frames (ORFs) present, again compared to less than one percent of the *E. coli* MG1655 genome. In concert with these differences it has been noted that there is an accelerated rate of gene loss in the *Shigellae* (Hershberg, Tang et al. 2007) and that there is also a reduced rate of gene acquisition and a lower probability of retaining newly acquired genes (van Passel, Marri et al. 2008).

Feature	Genome					
	E. coli MG1655	*S.dysenteriae* Sd197	*S.flexneri* 2a301	*S.flexneri* 2457T	*S.boydii* Sb227	*S.sonnei* Ss046
Chromosome Size (bp)	4,639,675	4,369,232	4,607,203	4,599,354	4,519,823	4,825,265
# ORFs	4254	4557	4434	4456	4353	4434
# Pseudogenes	12	285	254	372	218	210
# IS Elements	44	623	314	280	403	394
% Coding Sequence	87.3	77.2	80.4	77.2	80.5	80.5
% G+C	50.79	51.25	50.89	50.91	51.21	51.01
Deletions (kbp)	-	955	639	709	746	518
Insertions (kbp)	-	411	444	479	441	490
Translocations / Inversions > 5kbp	-	43	13	15	23	11

Table 3.1.1a – Comparison of Genomic features from five *Shigella* strains and *E. coli* MG1655 adapted from Yang et al 2005. The bottom three rows being comparisons to *E. coli* MG1655

3.1.2 – Summary

Given the identification of the *Shigellae* as specialised clones of *E. coli*, which have only recently adopted their current pathogenic lifestyle, and the availability of genome sequences for several strains of *E. coli* including at least one representative for each of the *Shigellae*, genomic comparisons become an option. This permits analysis of the changes associated with this recent adaption, specifically those changes which are not necessarily evident in the gross structure of the genome or in the extant composition, but trends which are apparent in the patterns of change in the genome.

3.1.3 – Aims & Conclusions

Here I aim to examine the evolutionary trends in members of the traditional genus *Shigella* as compared to other members of the *E. coli*, via examination of the pattern of nucleotide differences observable within the orthologues common to all nine genomes in the dataset. I observe a general preponderance of more deleterious mutations within the *Shigella* genomes, and detect a weakened time dependant purging of these deleterious mutations. A notable exception to this is *Shigella sonnei*, which shows patterns more akin to those observed in *E. coli*.

3.2 – Phylogeny Construction and Testing

3.2.1 – Neighbour-Joining Tree & Bayesian Topology Confirmation

A neighbour-joining (N-J) tree was constructed using a single sequence for each genome – the concatenated alignment of 2098 common orthologues, totalling approximately 2.1Mbp per taxon (~50% of the genome). This was compared to a Bayesian consensus tree from 50 random 20kbp segments and another from 50 evenly distributed 20kbp segments.

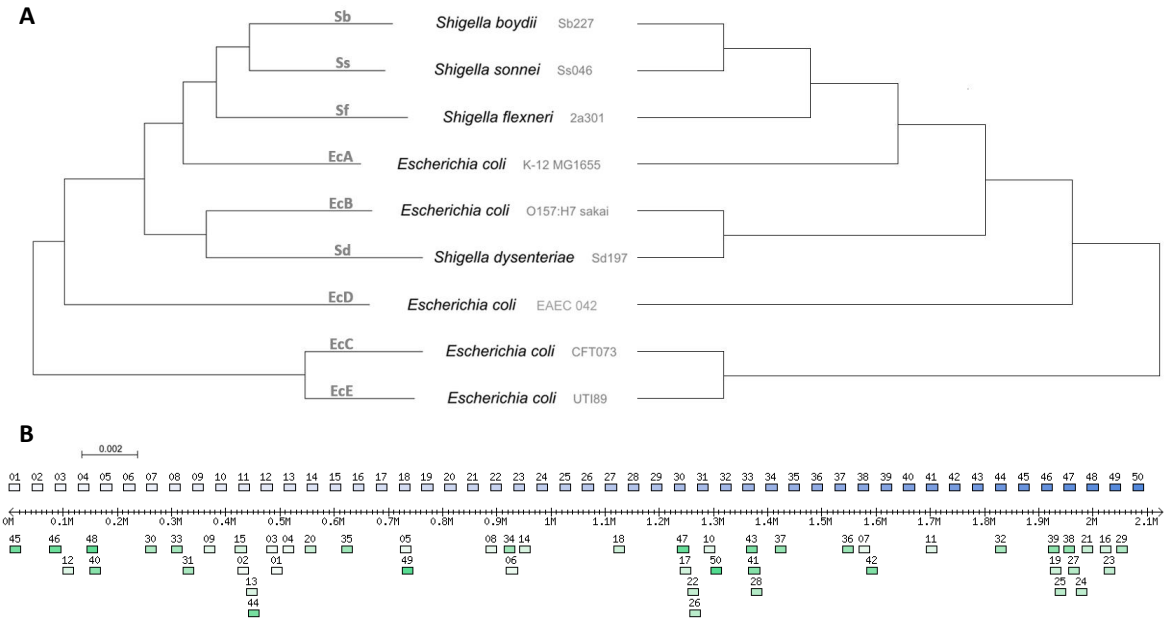


Figure 3.2.1a – A - Comparison of the N-J (left) and Consensus of Bayesian trees from random segments (right), the same topology was observed from the ordered segments. **B -** Comparison of the distributions of Bayesian segments for phylogenetic analysis. Evenly distributed above (blue) and randomly distributed below (green).

Given the agreement between the neighbour-joining and Bayesian topologies it can be inferred that the faster N-J algorithm is at least as accurate as a multi-sample Bayesian approach for sequences of this size. The bootstrap values for all nodes of the N-J tree were 100 percent, in line with the expected values given the large quantity of sequence data used.

3.2.2 – Comparison to Published Phylogeny

This phylogeny is in broad agreement with that from previous studies of the phylogenetic relationship between *E. coli* and the *Shigellae*, specifically compared with that by Yang et al (2007), one of the most comprehensive in terms of taxon use, there is only one disagreement; the Yang et al. phylogeny places *S. flexneri* as the outgroup to the clade of *S. boydii*, *S. sonnei* and *E. coli* K-12, whereas the phylogeny above places *E. coli* K-12 as the outgroup to an all-Shigella clade. However in a later phylogeny by van Passel et al (2008), shown in figure 3.1.1a, the taxa are positioned as in the phylogeny observed.

This rearrangement is relatively minor and could well correspond to the differences in data used when constructing the tree. Yang et al. used a total of 20,116bp of nucleotide sequence data for each taxon, providing 742 informative sites (~3.7%), the N-J phylogeny in figure 3.2.1a used ~2.1Mbp per taxon, providing 62,486 parsimoniously informative sites (~3.0%). Given that the van Passel et al phylogeny also used a larger quantity of sequence data (169 genes); it is possible that the phylogeny differences are due to the quantity of sequence data. However it is also important to note that the Yang et al phylogeny is comprised of 53 taxa, versus 14 in the van Passel phylogeny and 9 in the observed phylogeny, so it is likely that the presence of additional taxa reveals details of the relationships between taxa that would not otherwise be evident.

3.3 – Initial Analysis

3.3.1 – Internal Branch Method Comparison

The inclusion of internal branches to the analysis process maximises the information regarding the shared evolutionary history of the taxa involved and provides increased statistical support. The pattern of SNPs identified in the internal branches is independent of the patterns associated with terminal branches as all SNPs shared by a given clade are excluded during the initial SNP identification process.

The terminal branch subtraction (TBS) approach to estimating the pattern of SNPs associated with internal branches requires that each internal branch is compared with at least two “outgroup” taxa, the basic principal of the SNP direction assignment process. Based upon this criterion there are five suitable internal branches in the tree of this dataset, which are designated iA through iE and shown on the tree below (fig. 3.3.1a)

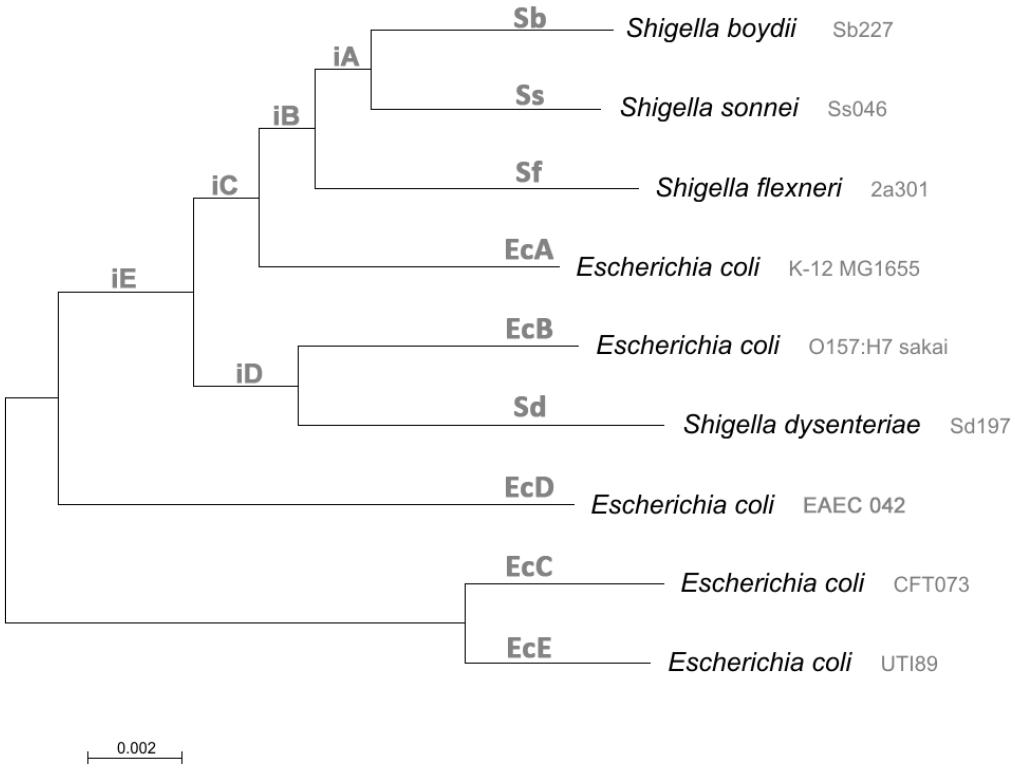


Figure 3.3.1a – Neighbour-Joining Tree showing the location of each of the five internal branches used.

Confirmation of the results from the TBS approach was performed by comparison to ancestral sequences inferred using the maximum likelihood approach as implemented in

PAML. The results (in terms of number of SNPs at each of the class of site) are listed below, with both TBS and PAML results shown side by side (table 3.3.1a).

Method	Internal Branch	Number of SNPs		
		NQ Sites	Q Sites	All Sites
PAML	iA	987	818	1805
	iB	636	692	1328
	iC	1268	1613	2881
	iD	1180	1406	2586
	iE	3441	4807	8248
TBS	iA	992	828	1819
	iB	1260	1312	2573
	iC	1533	1949	3481
	iD	1196	1449	2645
	iE	4543	6281	10826

Table 3.3.1a – The total number of SNPs at fourfold degenerate (Q) and non-fourfold degenerate (NQ) sites for each internal branch under both methods. The calculated values are the mean of all possible ways of estimating the branch values.

Whilst it is clear from the above table that the results from the two different methods are more similar for some of the internal branches than others (the most accurate being the two first-level internal branches iA and iD), the two sets of data correlate very strongly, showing correlation coefficients greater than 0.97 and p values less than 0.005. There is however a clear difference between the total number of SNPs identified at each site for the iB internal branch, there is no obvious explanation as to why this is the case but the proportional distribution of SNPs is similar between the two methods (52.1% at Q with PAML and 51.1 % Q with TBS).

It is notable that the TBS method produces consistently higher values than those identified from PAML ancestral sequences, a discrepancy which increases with the phylogenetic depth of the internal branch. It isn't obvious whether this is an overestimation in the TBS method or an underestimation of polymorphic sites by PAML. TBS overestimation can be considered to be less likely as TBS is based solely on the information derived from the extant sequence data, and assumes that any SNP common to any given group of terminal branches occurred in the common ancestor of all those taxa, in a parsimonious approach. In contrast the PAML derived ancestral sequences rely on inference of ancestral bases, which in cases of ambiguity may be excluded or the 'most likely' base chosen, these bases are then in turn used to infer SNPs associated with the internal branches, given a greater chain of assumptions between sequence data and SNP identification.

Method	Internal Branch	Ratio from Normalised SNP Counts	
		Tv / Ti	+AT / +GC
PAML	iA	0.4224	2.2991
	iB	0.3368	2.0387
	iC	0.3126	1.9093
	iD	0.3301	1.8403
	iE	0.2887	1.2218
TBS	iA	0.4221	2.3261
	iB	0.3606	2.0640
	iC	0.3222	1.7586
	iD	0.3356	1.8468
	iE	0.3110	1.2345

Table 3.3.1b – Comparison of Ratio values from normalised SNP counts from both methods. Tv/Ti = Transitions over Transverions and +AT / +GC = AT enriching over GC enriching SNPs.

Despite the differences in absolute numbers of SNPs identified at each nucleotide site type the relative proportions of each of the twelve directional SNP types (after normalisation) and ratios between specific groups of SNPs, such as the Transversion / Transtition ratio (Tv/Ti), show much greater agreement between the two methods (table 3.3b, above). So whilst the methods produce differing absolute values for SNP identification, they both capture the same inherent biases or trends with respect to the distribution of SNP types.

3.3.2 – SNP Site Distribution

Initial values of the total number of each type of SNP are uninformative; however analysis of the total number of SNPs at any given type of site, e.g. fourfold degenerate (Q), can potentially yield insights.

Species / Strain Code	Number of SNPs		
	NQ Sites	Q Sites	All Sites
EcA	2450	2100	4550
EcB	2967	2551	5518
EcC	2001	2598	5599
EcD	6261	6935	13196
EcE	2643	2437	5080
Sb	2636	1549	4185
Sd	5219	3397	8688
Sf	3628	2200	8528
Ss	2138	1477	3615
iA	992	828	1819 (4810)
iB	1260	1312	2573 (7042)
iC	1533	1949	3481 (9125)
iD	1196	1449	2645 (8426)
iE	4543	6281	10826 (15906)

Table 3.3.2a – The total number of SNPs. NQ & Q are non-fourfold degenerate (Non-Quartet) and fourfold degenerate (Quartet) sites respectively. 'All Sites' values for internal branches, bracketed in grey, represent the divergence "Time" of the SNPs, with the values for internal branches including only SNPs unique to the internal branch.

Looking at the proportion of SNPs occurring at the 100% synonymous fourfold degenerate ('quartet' or Q) sites, it is apparent that these largely represent 35-50% of the total number of SNPs (excluding internal branches). Although they represent under half of the SNPs it is important to note that the expected percentage of SNPs at Q sites is approximately 17%, reflecting that fourfold sites account for only 32 of the 192 possible sites in the genetic code (64 codons x 3 sites per codon). So the proportion of Q sites observed is far in excess of the expected mutational equilibrium, suggesting that selection has enriched the sequences for SNPs at Q sites.

3.3.3 – SNP Site Distribution over Time

Both of the patterns observed above are simply extant features of each taxon, given the variation in branch lengths and therefore associated divergence time, it is possible to examine the trend of these patterns with respect to time for the set of genomes as a whole. Initial analysis of the trends observed in proportion of SNPs occurring at Q sites, reveals a trend which is congruent with known evolutionary principles as all Q site SNPs are completely synonymous, the gradual purging of the nonsynonymous NQ site SNPs results in the proportion of the observed SNPs at Q sites increasing with divergence time, this effect can be seen in figure 3.3.3a. The trend with time is well supported statistically, with the p value ($p < 0.001$) and also an adjusted R^2 value of 0.631.

It is also apparent from the trend observed that selection has not acted instantaneously and has taken time to preferentially purge certain SNPs, with 'older' SNP patterns showing a greater bias in the location of SNPs than 'younger' counterparts. Given the relatively close relationship (i.e. low divergence times) of these taxa it is perhaps unsurprising to find that the selective purging of deleterious mutations is not complete, given the time dependence of the process (Rocha, Maynard Smith et al. 2006).

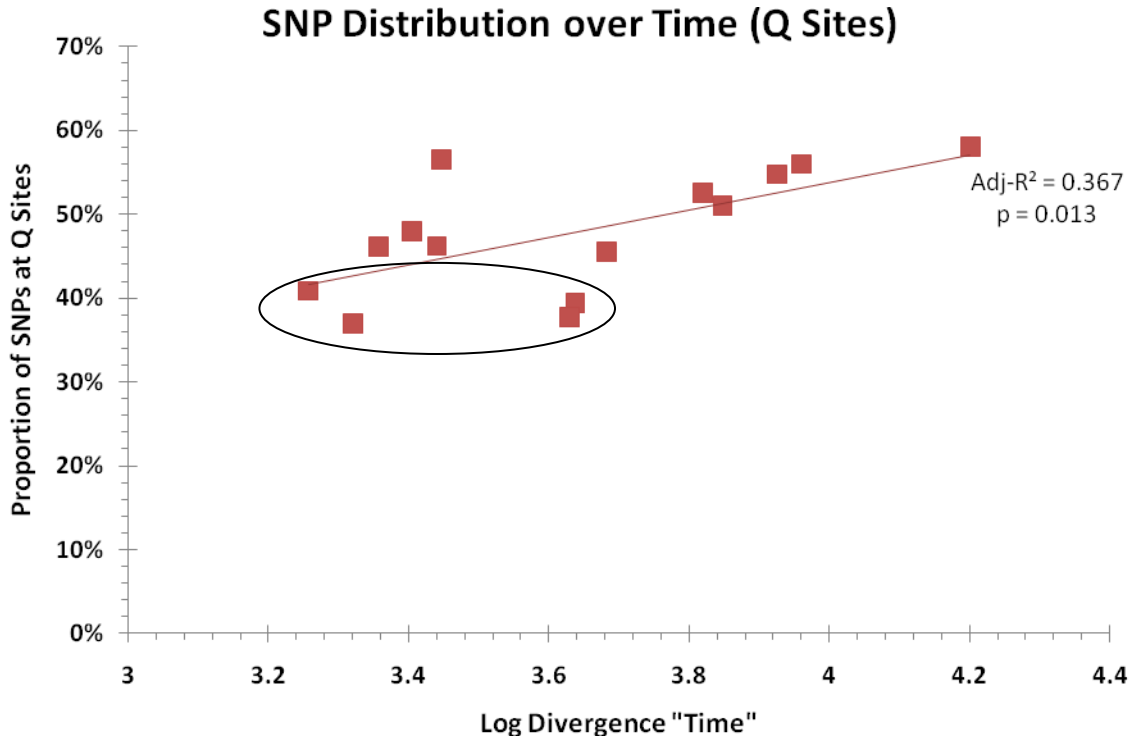


Figure 3.3.3a – The proportion of SNPs at fourfold degenerate positions against divergence "time" for each taxon, showing a linear least-squares regression line with Adjusted R^2 and p values. The ringed points are the *Shigellae*.

The exception to the general trend for the data appears to be the *Shigellae*, which show a much slower trend of enrichment for more synonymous SNP locations over time than *E. coli* (regression lines based on the *Shigellae* alone show no significant trend – $R^2 \approx 0$ and $p > 0.10$); as well as having a higher proportion of SNPs at nonsynonymous sites. This hints at a possibility of differing selection pressures and/or modes of selection acting upon the *E. coli* than the *Shigellae*.

3.4 – d*N*/d*S* Ratio Analysis

3.4.1 – d*N*/d*S* Ratio of the observed SNPs

In light of the observed enrichment for more nonsynonymous SNP sites over time, an examination of the relative rates of synonymous and nonsynonymous substitution within these changes becomes the next logical step.

Species / Strain / ID	d*N*/d*S* Ratio	Log Divergence "Time"
EcA	0.1069	3.3570
EcB	0.1102	3.4408
EcC	0.1063	3.4472
EcD	0.0559	3.8194
EcE	0.0902	3.4048
Sb	0.2489	3.3208
Sd	0.2185	3.6379
Sf	0.2531	3.4645
Ss	0.1921	3.2572
iA	0.1069	3.6821
iB	0.0524	3.8477
iC	0.0442	3.9602
iD	0.0447	3.9256
iE	0.0339	4.2016

Table 3.4.1a – The d*N*/d*S* ratio of the SNPs within a taxon, and the associated Divergence "Time"

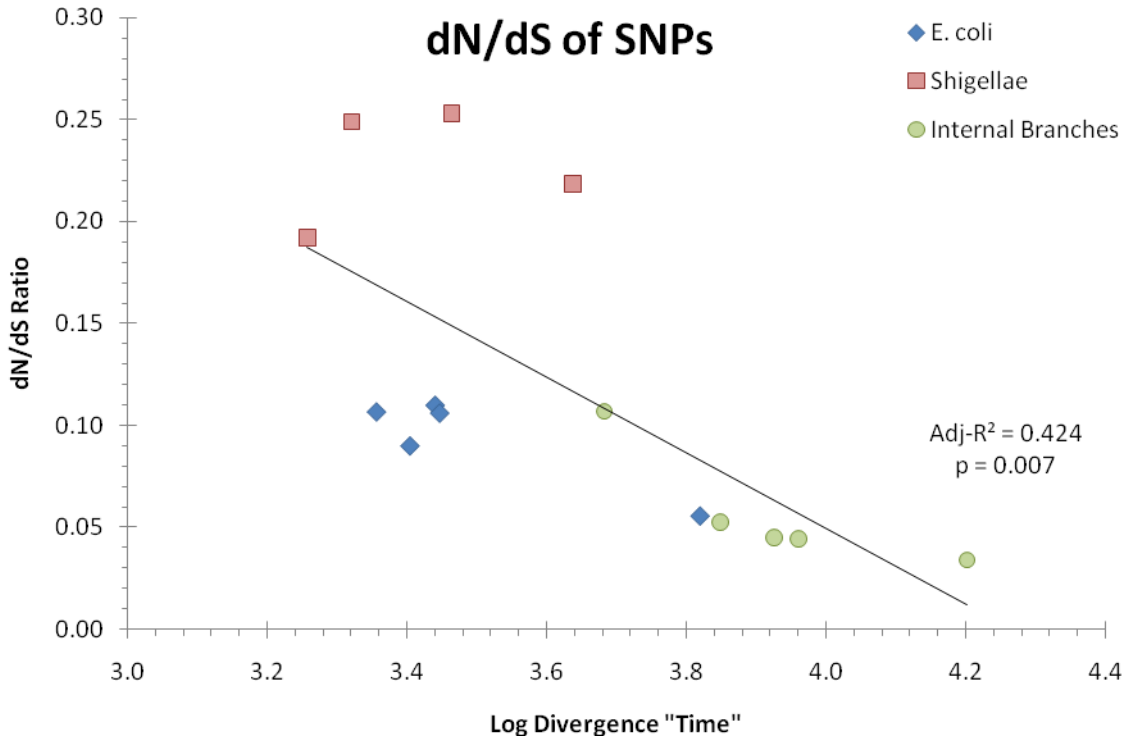


Figure 3.4.1a – The d*N*/d*S* ratio plotted against Log Divergence "Time" associated with the SNPs from which the ratio is derived.

As originally observed in the change in the distribution of the SNPs with respect to “time” the d*N*/d*S* ratio shifts towards favouring synonymous changes (d*S*) with greater divergence times, reflecting the time delayed expunging of the largely deleterious nonsynonymous changes. This trend, whilst not as strong (Adj-R^2 of 0.424) as the trend towards Q sites or indeed 3rd sites (see Chapter 4), is still highly significant, with a p value of 0.007. This trend is a stronger indication than the corresponding trends observed (Ch4) as the process of calculating the d*N*/d*S* ratio takes into account the rates of nonsynonymous and synonymous changes per corresponding site, converting each absolute count of SNPs into a proportion of the possible sites of that type. Thus there is an absolute distinction between the synonymous and nonsynonymous changes, compared to the ‘more synonymous’ or ‘less synonymous’ qualifiers associated with using either Q sites or 3rd codon positions.

Again with this trend there is a distinct difference between the behaviour of *Shigellae* and *E. coli* with the former showing a propensity towards a higher d*N*/d*S* ratio, and lack of a trend in the proportion of synonymous SNPs with time (Adj-$R^2 \approx 0$). This adds to the strong contrast in the differing selective pressures that are acting or have been acting upon the *Shigellae* than the other *E. coli*.

Given the lifestyle of the *Shigellae* it is likely that their intracellular mode of replication and proliferation during their infection of a human host has resulted in these differential selective pressures, in contrast with the pathogenic *E. coli*, which occupy a similar niche but do not replicate intracellularly.

3.4.2 – d*N*/d*S* Differences within Functional Classes of Gene

Whilst the data above agrees with the trends observed by Rocha et al (2006) and strongly suggests reduced purifying selection as the cause of the differences observed between *Shigellae* and *E. coli* it is also possible that positive selection may be responsible.

However, there is a largely consistent difference between the *Shigellae* and the *E. coli* in each of the gene categories examined (figure 3.4.2a), in line with the surfeit of nonsynonymous polymorphisms observed in the *Shigellae* across all orthologues

examined above. This consistent signature of more rapid evolution in *Shigellae*, along with the lack of any signature of an even greater rate of evolution in genes more likely be under host immune attack (specifically, Cell Envelope genes), renders positive selection an unlikely cause of the observed differences.

The majority of the functional categories also display d*N*/d*S* ratio differences which are greater than that which would be observed from a random comparison of the average of any 5 versus the average of the remaining 4 genomes from the dataset (reflecting the 5 *E. coli* and 4 *Shigellae*), this illustrates that the observed bias is not an effect of the unequal number of taxa in each group. The one exception to this is the genes associated with transcription, which show no greater difference between the *Shigellae* and *E. coli* than would be expected from any random comparison, possibly reflecting the relatively high selective constraint that a crucial class of gene would experience, mutations in any other given category may disable a single gene or reduce its efficiency, mutations to transcription genes disable or seriously hamper the expression of all genes, carrying a significantly higher selective cost.

One other striking class of gene is fat metabolism; the surfeit of nonsynonymous polymorphisms in these genes is indicative that some aspect of the difference between *Shigellae* and *E. coli* has resulted in the rapid evolution of these genes relative to the rest of the genome.

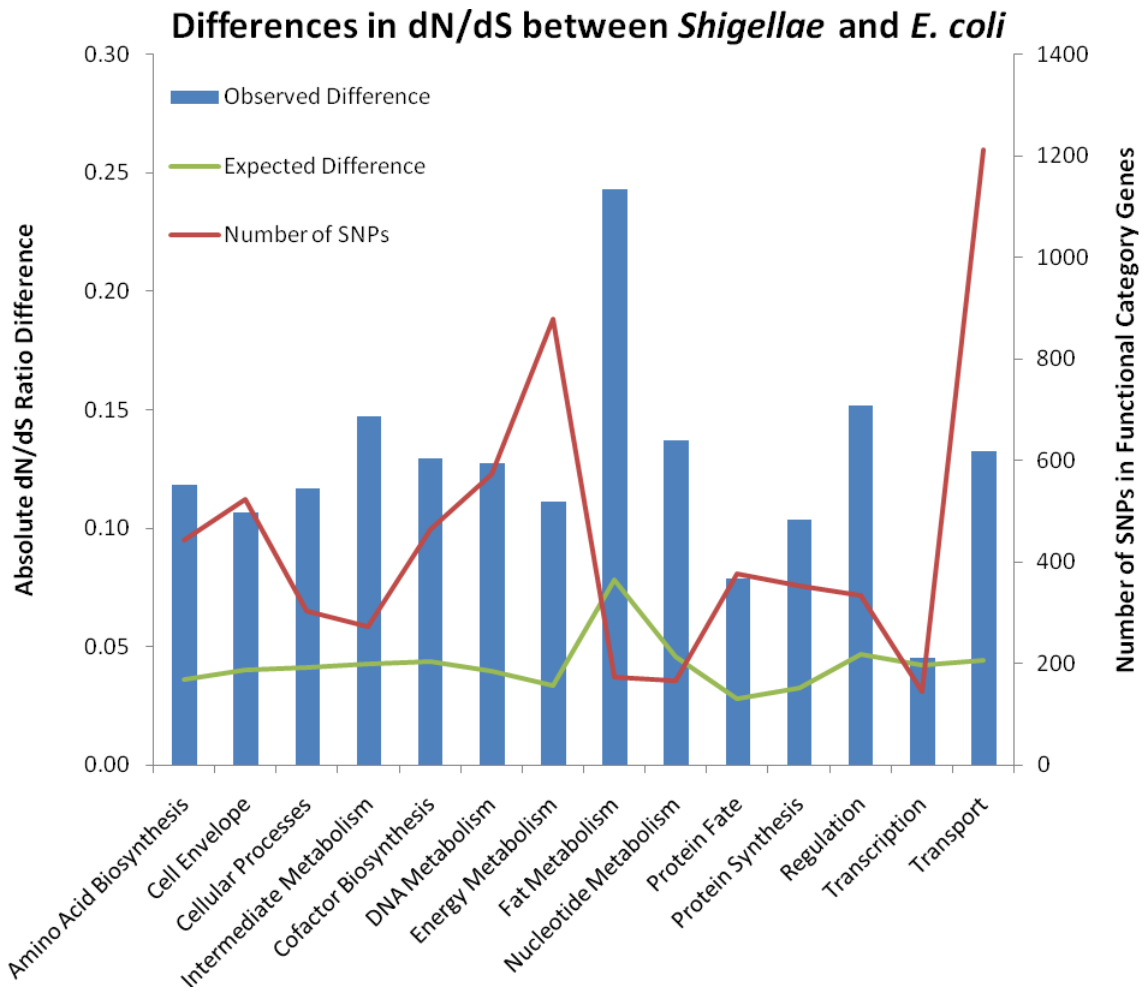


Figure 3.4.2a – The absolute d*N*/d*S* ratio difference between *E.coli* and *Shigellae* for various functional categories of gene. The expected difference is based upon the mean of all possible comparisons of 5 ratios versus 4 ratios, reflecting the 5 *E. coli* and 4 *Shigellae*.

3.5 – Principal Component Analysis

As per the methods, the polymorphism profiles of each of the terminal and internal branches were used as input data for the Principal Component Analysis (PCA). The data was input as 12 variables, each one representing the polymorphism profile values of a given SNP type. The analysis recovered 12 principal components explaining the data and their associated weightings and statistics.

3.5.1 – Trends in Principal Components

Of the twelve principal components recovered from the analysis, only three explain a significantly greater proportion of the variation than an input variable (i.e. greater than 1/12th or 8.333% of the variation observed); principal components (PCs) one through three explaining 43.4% 20.3% and 15.6% respectively, these values are derived from the

eigenvalues associated with each principal component, which in themselves also represent a statistical test of the level of variation explained by a principal component, values less than 1 representing a loss of information content relative to the input variables, see table 3.5.1a below.

Principal Component (PC)	Eigenvalue	Variation Explained (%)
1	5.206	43.4
2	2.440	20.3
3	1.866	15.6
4	1.006	8.4
5	0.442	3.7
6	0.401	3.3
7	0.361	3.0
8	0.107	0.9
9	0.099	0.8
10	0.047	0.4
11	0.020	0.2
12	0.005	0.0

Table 3.5.1a – The Eigenvalues and percent of variation explained by each principal component

Plots of the scores for each taxon/internal branch within, each of the first three, principal components against time (figures 3.5.1a & b, below) shows that only the first principal component shows the trends observed in the nucleotide distribution and d*N*/d*S* ratio what's more this trend is significant and strong (Adj-R^2 = 0.751 and $p < 0.001$), notably the *Shigellae* are showing slightly higher scores than the *E. coli* as seen in the nucleotide analysis trends. The second and third principal components show no significant trend with divergence time; which is expected given they will show no significant correlation to PC1 and that correlates with 'time' strongly.

This lends further credance to the differences observed previously between *E. coli* and *Shigellae*, given its emergence here as a trend present in the values of a principal component explaining 43% of the variation in the data.

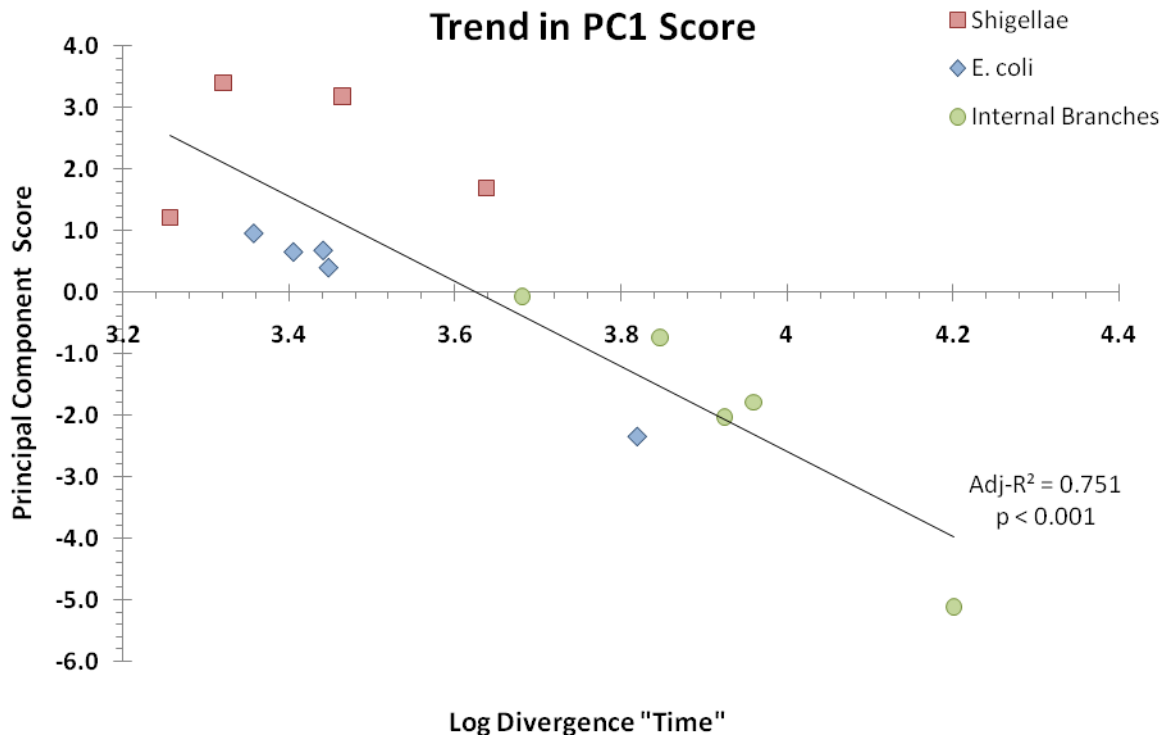


Figure 3.5.1a –The Principal Component 1 score plotted against divergence time for each taxon / internal branch.

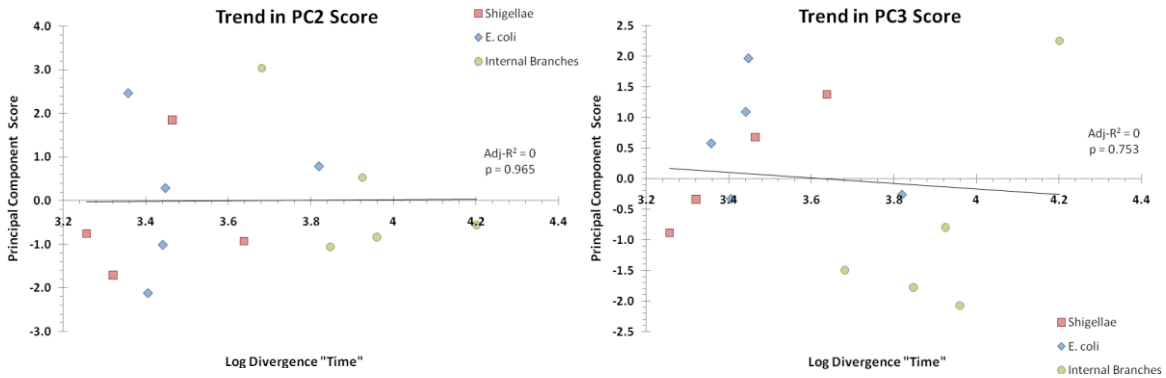


Figure 3.5.1b – The Principal Component 2 & 3 scores plotted against divergence time for each taxon / internal branch, showing no correlation with time as expected.

3.5.2 – Differences between *Shigellae* and *E. coli*

Full confirmation of the differences observed can be seen in figure 3.5.2a (below); comparison of the actual difference observed to the 'expected' difference (the mean of all possible combinations of 'mean of 5' minus 'mean of 4', reflecting the 5 *E. coli* versus the 4 *Shigellae*) in figure 3.5.2b shows that of all the principal components it is only principal component one that simultaneously explains a large share of the variation of the dataset and shows a much higher observed difference than would be expected by random chance, the observed difference is actually the maximum possible difference (excluding the internal branch values).

Principal components 2 through 7 (boxed in figure 3.5.2b) show either no greater difference than expected by random chance (PCs 2,3 & 4) or explain an insignificant amount of variation to be informative (PCs 5, 6 & 7). Principal components 8 through 12 all both explain an insignificant proportion of the variation and show no appreciable difference between the observed and expected values.

Whilst this firmly supports the differences between the *Shigellae* and *E. coli* it also implicates them as the major source of variation in the data, given the principal component displaying them is the only significant one and explains 43% of the total variation.

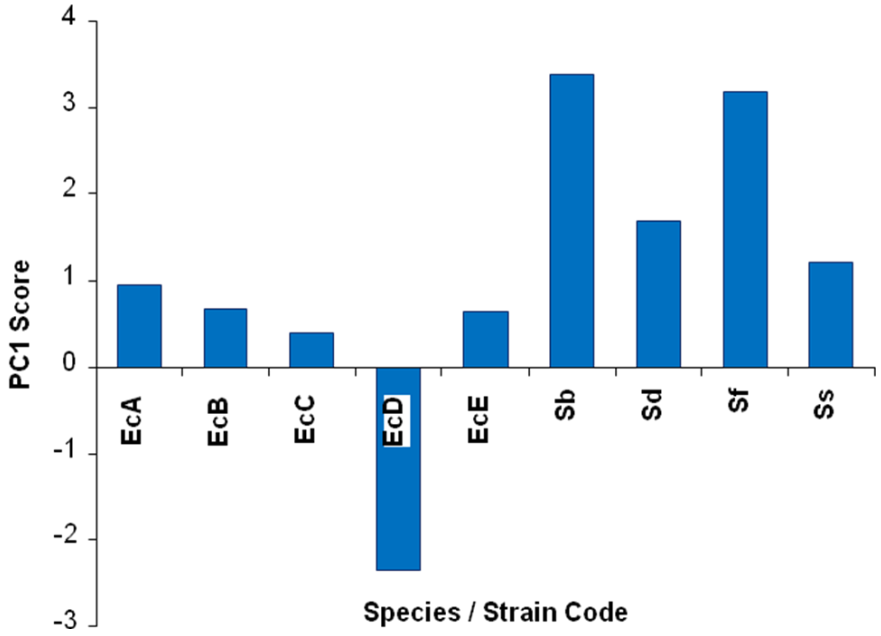


Figure 3.5.2a – The PC 1 scores for each taxon.

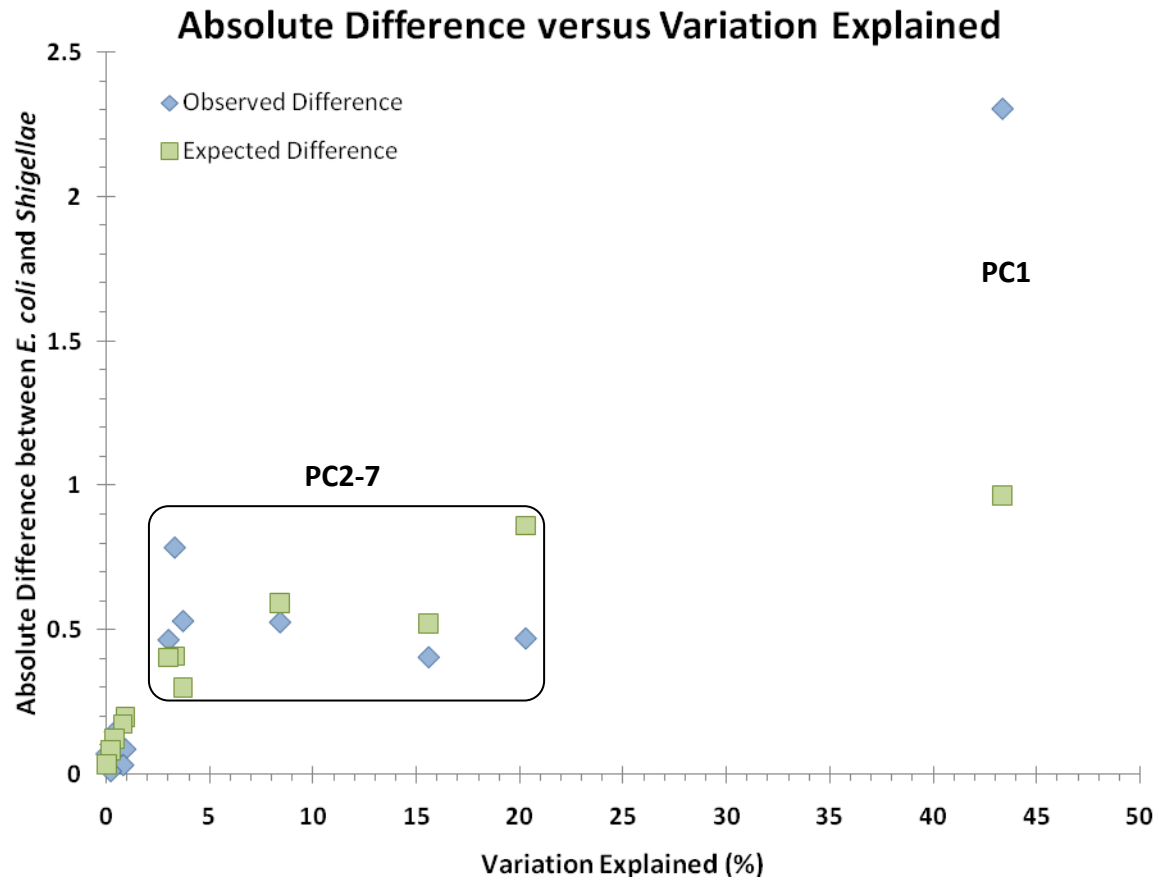


Figure 3.5.2b – The absolute difference between *Shigellae* and *E. coli* plotted against the percent variation explained for each of the Principal Components. PCs 2 through 7 are boxed.

3.5.3 – Principal component breakdown

Whilst the first principal component isolates the primary trend responsible for the variation in the data, the principal component scores have no biological meaning associated with them, which obfuscates interpretation of the data. Deriving biologically meaningful information from the principal component requires examination of the relative weightings of the input variables. The patterns observed in the principal component scores will be reflected by the ratios of the most positively weighted and most negatively weighted input variables.

Using the weightings from PC1 and dividing the input variables (SNP types) by two different categories (table 3.5.3a), enables the description of biological effects associated with the trend identified by PCA. The effect of the different SNP types on genome base composition – either GC enriching or AT enriching (+GC or +AT respectively) fits quite closely with the weightings, the top four most positively weighted SNP types are all AT enriching, whilst the four most negatively weighted comprise two GC enriching and two

enrichment neutral SNP types. In total the +AT SNPs have a weighting of +1.42 and the +GC SNPs have a weighting of -0.65, suggesting that the ratio of AT enriching to GC enriching SNPs is an explanation of the variation in the data.

Less strongly, there is a similar bias in the total weightings for Transversions (Tv) and Transitions (Ti) (+0.97 and -0.13 respectively). The transitions and transversions aren't as clearly biased towards one extreme or the other as the +AT/+GC SNP types, however given that their summed weightings still show a total positive - negative split, they may also be indicative of the variation explained by PC1.

Input Variable	PC1 Weighting	Nucleotide Enrichment	Transition (Ti) Transversion (Tv)
G→A	0.380	+AT	Ti
C→A	0.370	+AT	Tv
G→T	0.356	+AT	Tv
C→T	0.315	+AT	Ti
A→T	0.225	-	Tv
A→C	0.147	+GC	Tv
C→G	0.140	-	Tv
T→G	0.030	+GC	Tv
G→C	-0.061	-	Tv
T→A	-0.235	-	Tv
A→G	-0.398	+GC	Ti
T→C	-0.428	+GC	Ti

Table 3.5.3a – Table of the weighting of the twelve input variables for PC 1, as well as the effect of the SNP type on genome base composition (AT or GC enriching) and type of the SNP in terms of transition or transversion.

Overall the first principal component, explaining 43% of the observed variation in the data, supports a clear difference between the *E. coli* and the *Shigellae* in line with previously observed trends, furthermore the weightings of the input variables within PC1 indicate that the trends described by PC1 are changes in the nucleotide bias of the SNPs and the transition / transversion bias of the SNPs.

3.6 – Metric Ratio Analysis

3.6.1 – Choice and Calculation of Metric Ratios

Based on the results from the principal component analysis, the metric ratios of +AT/+GC and Ti/Tv were chosen and calculated using the normalised polymorphism profiles for each of the 9 taxa and 5 internal branches, at all sites and at both Q and NQ sites. The ratios are listed below (Table 3.6.1a & b for +AT/+GC & Ti/Tv respectively).

Taxon / Branch ID	Divergence Time	Ratio of AT enriching to GC enriching SNPs (+AT / +GC)		
		All Sites	NQ Sites	Q Sites
EcA	3.357	2.320	2.282	1.627
EcB	3.441	2.348	2.318	1.659
EcC	3.447	2.179	2.191	1.529
EcD	3.819	1.708	1.714	1.127
EcE	3.405	2.412	2.442	1.675
Sb	3.321	2.878	2.875	2.219
Sd	3.638	2.485	2.467	1.820
Sf	3.464	2.871	2.792	2.311
Ss	3.257	2.677	2.621	2.021
iA	3.682	2.326	2.256	1.698
iB	3.847	2.064	1.997	1.466
iC	3.960	1.759	1.806	1.128
iD	3.926	1.847	1.784	1.260
iE	4.201	1.235	1.272	0.774

Table 3.6.1a – The +AT/+GC ratio for each Taxon and internal branch (estimated using TBS) for All sites, NQ and Q sites.

Taxon / Branch ID	Divergence Time	Transition/Transversion Ratio (Ti / Tv)		
		All Sites	NQ Sites	Q Sites
EcA	3.357	2.518	3.325	1.735
EcB	3.441	2.888	3.842	2.049
EcC	3.447	2.811	3.635	2.018
EcD	3.819	2.978	5.221	1.871
EcE	3.405	3.291	4.824	2.251
Sb	3.321	2.370	2.683	1.824
Sd	3.638	2.707	3.086	2.138
Sf	3.464	2.276	2.348	1.955
Ss	3.257	2.949	3.453	2.289
iA	3.682	2.369	2.724	1.917
iB	3.847	2.773	4.742	1.714
iC	3.960	3.103	6.008	1.933
iD	3.926	2.980	5.457	1.861
iE	4.201	3.215	6.880	2.102

Table 3.6.1b – The Ti/Tv ratio for each Taxon and internal branch (estimated using TBS) for All sites, NQ and Q sites.

The ratio of AT enriching to GC enriching SNPs is consistently greater than one (with the singular exception of the phylogenetically deepest internal branch – iE, at Q sites), this reflects the relative abundance of AT enriching SNPs, displaying an expected mutational bias congruent with the established GC bias in the genome composition, which stands at approximately 51% GC in the genome as a whole or closer to 53% GC in the "core genome" under study. In addition the predominant mutation incurred by a genome is C→T a result of the deamination of Methyl-cytosine; the resulting abnormal base is recognised as a thymine by the DNA repair and replication machinery.

The +AT/+GC ratio observed is not significantly different between 'All' sites and NQ sites; a paired t-test gives a p value of 0.453. There is however an obvious and significant difference between the NQ and Q sites, with the latter being consistently lower (mean differences of 0.021 and 0.607 for All versus NQ and Q versus NQ respectively). This is possibly a selective bias as the favoured codons used by *E. coli* and *Shigellae* are GC rich at Q sites and so AT enriching mutations at Q sites would carry a relatively higher selective cost than their counterparts at NQ sites where codon bias is less affected by GC content.

The ratio of Transitions (Ti) to Transversions (Tv) is significantly greater than one, reflecting the expected bias towards transitions given the more deleterious nature of a nonsynonymous transversion as compared to a nonsynonymous transition (Zhang 2000). There is also a mutation bias in favour of transitions, partly due to the C→T mutation predominance mentioned earlier, and documented in *Caenorhabditis elegans* (Denver, Morris et al. 2004).

The Ti/Tv ratio is significantly higher at NQ sites than both 'All' and Q sites, reflecting the absolute non-degeneracy of any transversions at these sites (the only degeneracy remaining – twofold sites are always linked via transitions at the 3rd codon position) so any selective bias would be heavily in favour of transitions. The values of the ratio at Q sites are all close to 2 reflecting the selective equality of transitions and transversions at these sites and so they largely reflect the 2:1 mutation bias.

3.6.2 – Metric Ratio Confidence

The graphs below (figures 3.6.2a & b) show both the +AT/+GC value and the Ti/Tv values for each taxon and internal branch at all nucleotide sites. The error bars are calculated from 1000 bootstrap replicates of the polymorphism profiles for each taxon; from each profile the metric ratios are calculated, these are then used to calculate the standard deviation of the replicate ratios. The error bars represent 1.96 standard deviations which is equivalent to 95% of the bootstrap replicates and is representative of a 95% chance that the 'true' metric ratio is in that interval.

As can be seen the 'confidence intervals' are relatively small indicating little error beyond the edges of the point markers on the graph. The larger intervals are associated with the internal branches, an expected result given that these points are inferred rather than directly measured, even so even the largest interval (Ti/Tv for iB) still does not even approach the spread of the observed points. Overall the observed ratios can be taken to be reliable estimators of the 'true' associated with the polymorphism profiles.

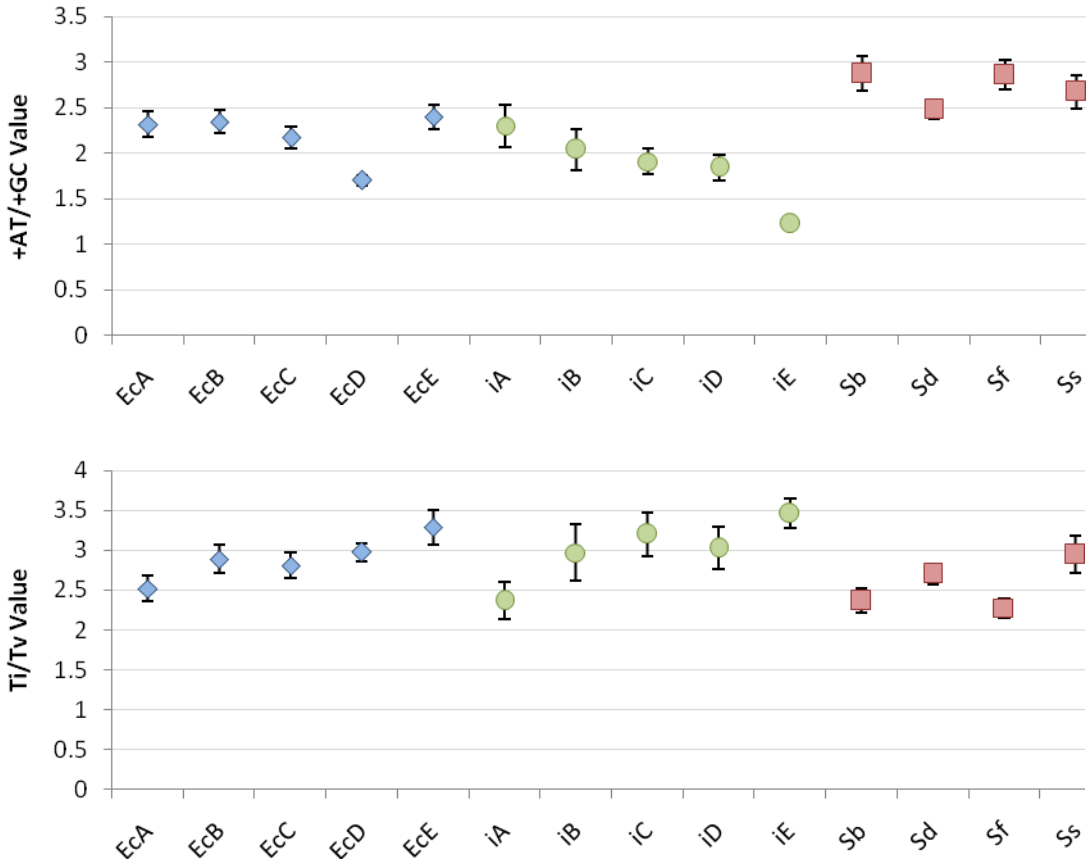


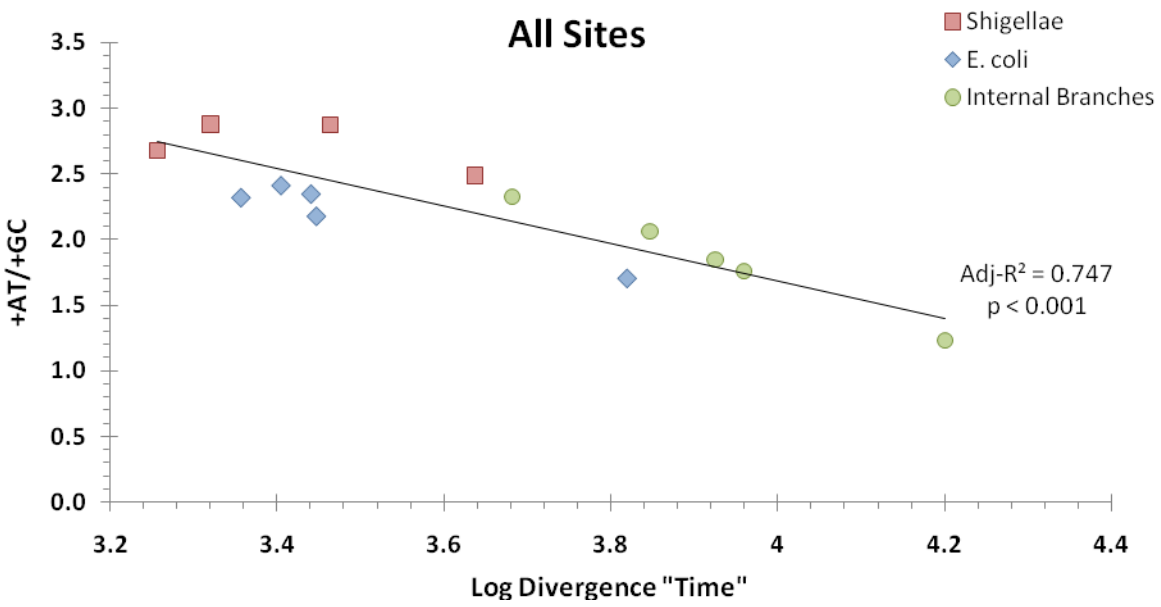
Figure 3.6.2a & b – +AT/+GC and Ti/Tv metric ratio values with confidence intervals of 1.96 standard deviations for each taxon, at 'all' nucleotide sites.

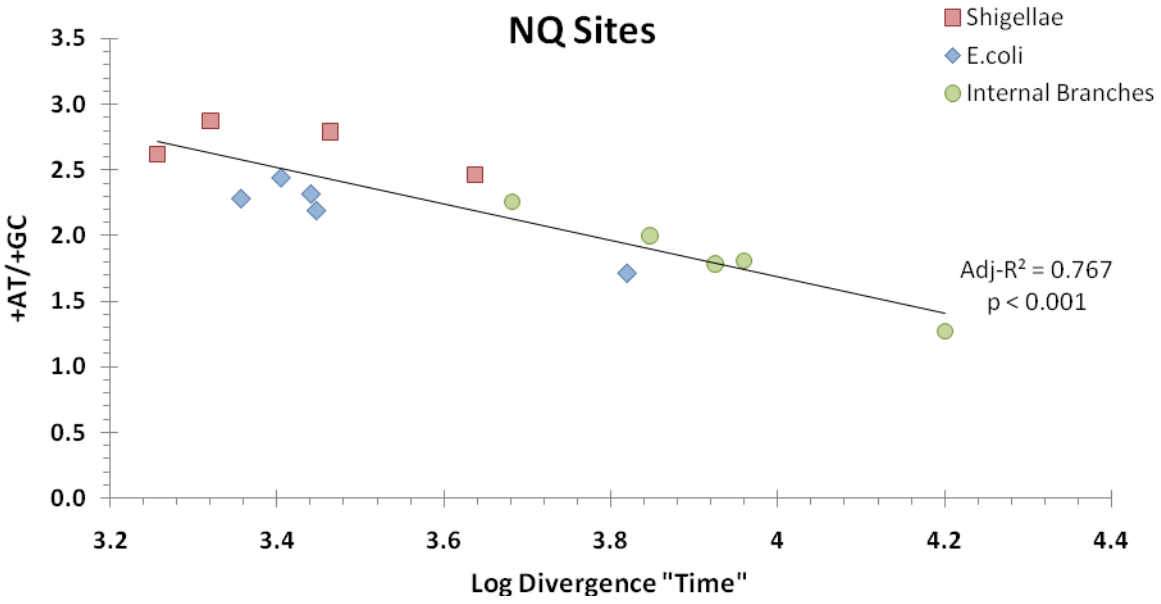
3.6.3 – AT versus GC enrichment over time

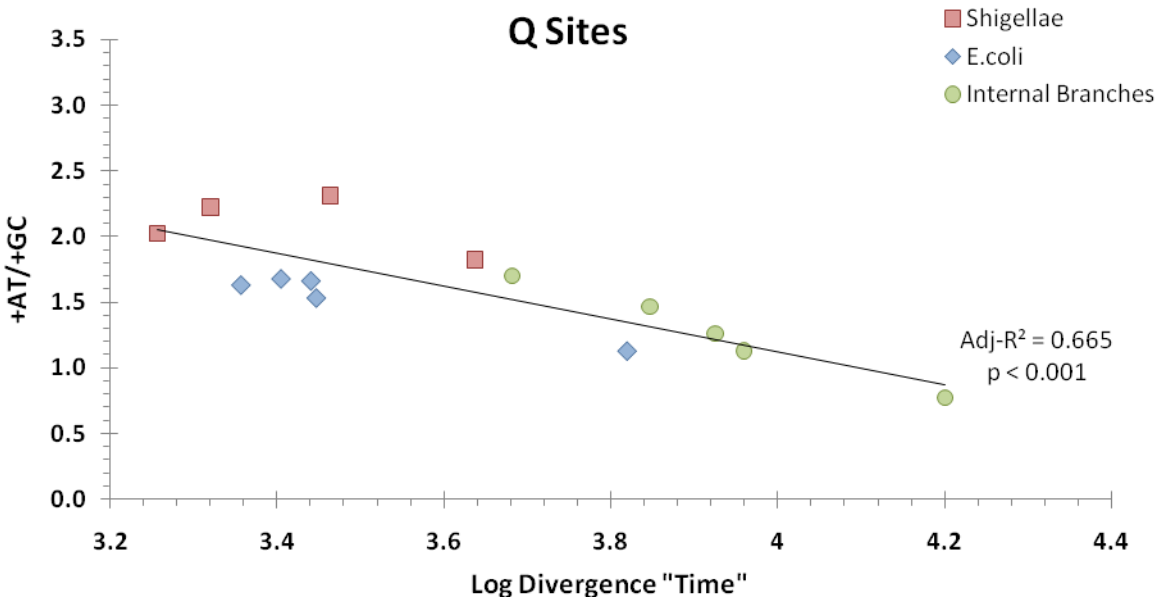
Regardless of the breakdown by class of nucleotide site, the generally observed trend towards gradual decrease of the +AT/+GC ratio holds (figure 3.6.3a,b&c). This gradual decrease correlates with the purging of AT enriching SNPs over evolutionary time, a situation expected as the more closely related species or strains would exhibit differences more akin the inherent mutational bias, which given the predominance of C→T mutations is towards +AT/+GC ratio values far greater than 1. Selection then acts in the longer term to purge the AT enriching SNPs, restoring the genome AT content to the selectional equilibrium, which manifests as a gradual decrease in the +AT/+GC ratio.

The *Shigellae* show a greater proportion of AT enriching SNPs than *E. coli*. That these differences hold across site types, the trend being nearly identical at both NQ and Q sites, even though the overall values of the ratio are somewhat lower at Q sites, strongly implies that the effects on the +AT/+GC ratio are not solely derived from effects associated with the encoded amino acid but include general features of DNA itself with implications for stability and structure of the physical genome.

S. sonnei shows a difference in behaviour from the other *Shigellae*, with a much lower ratio than the rest of the *Shigellae*, this difference is evident when considering the +AT/+GC ratio of the SNPs at all nucleotide sites and is present when examining the ratio at Q and NQ sites independently.







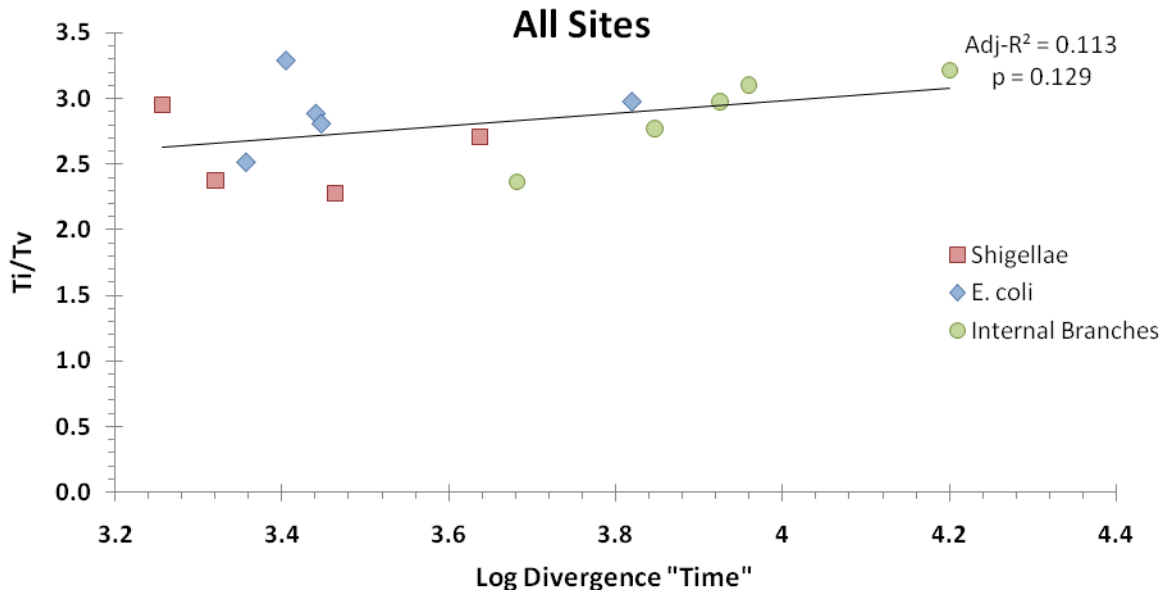
Figures 3.6.3 a, b & c – The +AT/+GC ratio versus time at all sites, NQ sites and Q sites respectively.

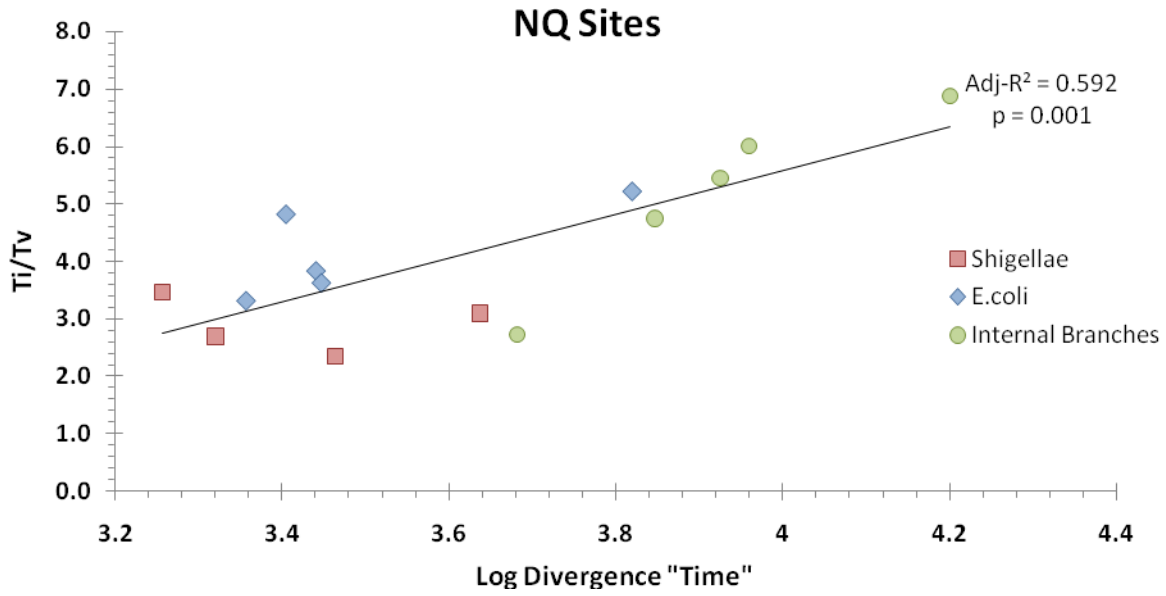
3.6.4 – Transitions versus Transversions over time

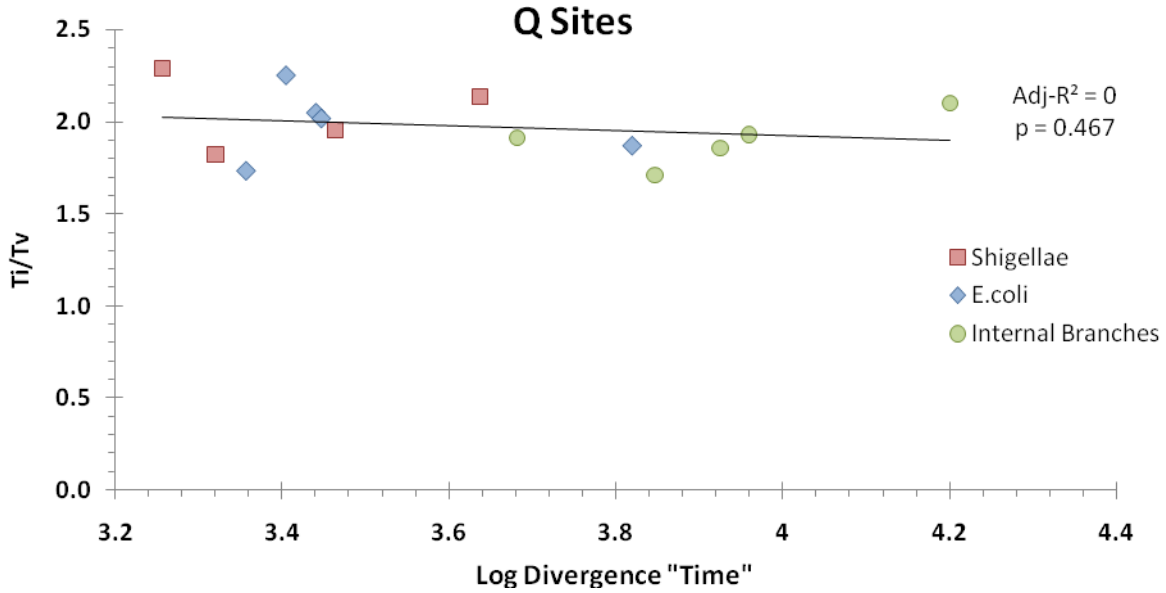
The ratio of transitions to transversions (Ti/Tv) shows a significant downward trend with divergence time when considering all sites and when considering NQ sites independently (figures 3.6.4a,b&c). The downward trend reflects the gradual purging of the more deleterious Transversions, which are selectively disfavoured as on average nonsynonymous transversions are less conservative than nonsynonymous transitions so incurring a greater selective cost.

Given the strong correlation to the inverse of the d*N*/d*S* ratio, it is not unsurprising that the Ti/Tv ratio is strongest when largely synonymous sites are excluded, as there is no significant difference on the selective impact of either transitions or transversions when there is no effect on the amino-acid encoded. This difference is clear when considering the trend with time at NQ sites (Adj-R^2 = 0.592, p = 0.001) versus that at Q sites (Adj-R^2 = 0, p = 0.467), where the former shows a strong relationship with time that is highly significant and the latter shows no significant relationship.

As with +AT/+GC, where there is a significant trend with time, there is a propensity for *S. sonnei* to exhibit more similar patterns to *E. coli* than the other *Shigellae*, in this case that is reflected in a slightly higher Ti/Tv ratio in *S. sonnei* than in the other *Shigellae*.







Figures 3.6.4 a, b & c – The Ti/Tv ratio versus time at all sites, NQ sites and Q sites respectively.

3.6.5 – Regression Residual Analysis

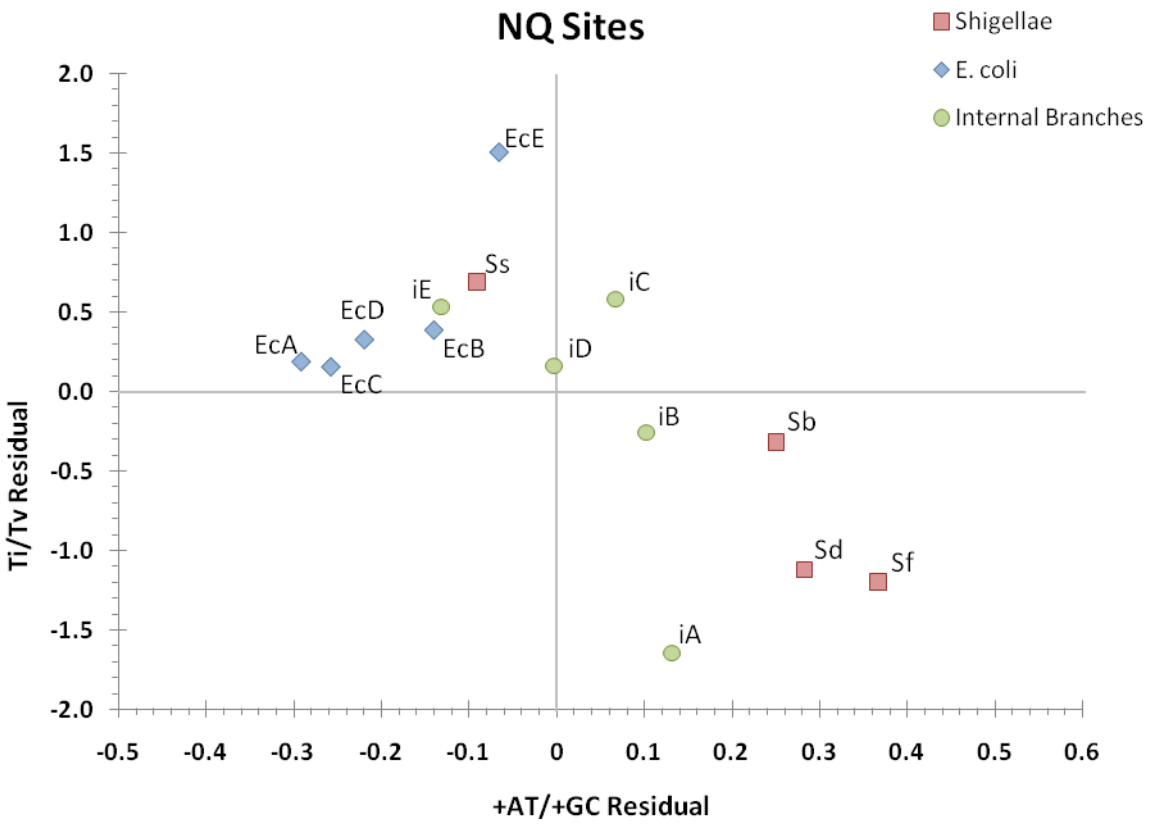
Using the Q and NQ site regressions, scatter plots of residuals at each site were constructed, with the +AT/+GC residual on the x-axis and Ti/Tv Residual on the y-axis, to examine the separation of the taxa and internal branches described by these metrics, whilst still correcting for the effects of divergence time.

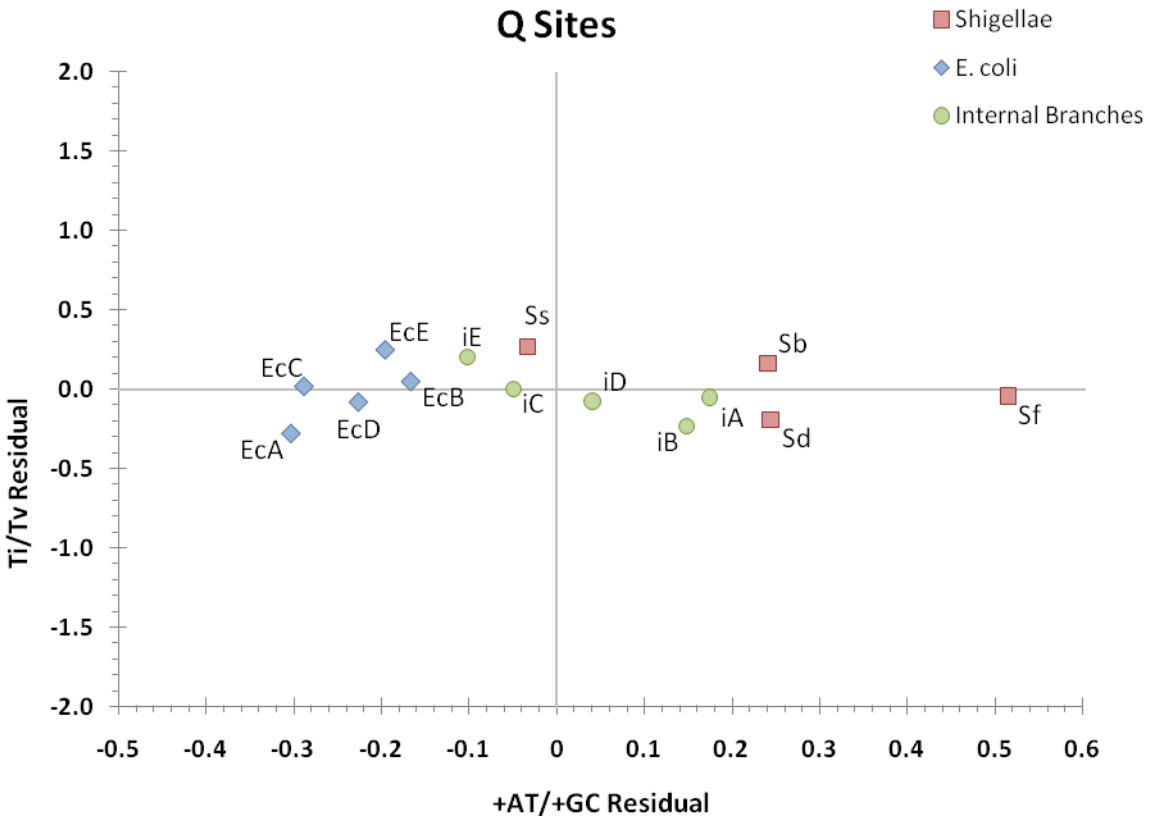
It is fairly evident from Figures 3.6.5 a & b that the TiTv residual at Q sites contributes no information to the spread of the data points, as they are all roughly equally distributed around zero. However there is clear pattern from the +AT/+GC residual (x-axis) on the Q site scatter plot (3.6.5b) and the NQ scatter plot (3.6.5a) which separates the majority of the *Shigellae* from the *E. coli* with the notable exception of *S. sonnei*. The internal branches forming an intermediate group between the two, with the branches supporting a majority of *Shigellae* being closer to the main *Shigella* cluster.

The presence of a *Shigella* cluster is likely a consequence of the differences in lifestyle between the *Shigellae* and the *E. coli* the former invading host cells and replicating intracellularly and the latter either occupying a more commensal niche or invading the gut lining but replicating in the extracellular space.

The residuals reflect the relative predominance of +AT SNPs over +GC SNPs in the *Shigellae* (as shown by their positive x-axis values), at both NQ and Q sites, as well as a relatively greater proportion of transversions (negative value on the y-axis), in this case only at NQ sites. In both cases these have resulted in a greater proportion of the more deleterious SNP type in *Shigellae*, indicating that they are perhaps experiencing different evolutionary effects to the *E. coli*.

As observed previously *S. sonnei* shows patterns much more consistent with the *E. coli*, this is possibly related to the relative youth of the *S. sonnei* lineage as compared to the other *Shigellae* (10,000 years versus 35,000 - 270,000).





Figures 3.6.5 a & b – Scatter plots of the residuals to the regression lines with respect to divergence time for +AT/+GC and Ti/Tv at NQ and Q sites respectively, all taxa / internal branches are labelled with their short codes.

3.6.6 – Bootstrap Confirmation of Differences

Based upon the percentile value results of bootstrapping the metric ratios calculated from polymorphism profiles from 'All Sites', there is a significant difference between the *Shigellae* and the *E. coli*, indicated by percentile values either greater than 95% or less than 5%. Where greater than 95% indicates that the *Shigella* is higher than the 95% of the bootstrapped *E. coli* replicates and less than 5% indicates that it is lower than 95% of the bootstrap replicates, it is important to note however that the bootstrap comparisons do not control for divergence time.

With the +AT/+GC ratio the *Shigellae* are consistently higher than the *E. coli*, showing percentile values of 100% across the board (table 3.6.6a). This supports the earlier observations that the *Shigellae* show a greater proportion of AT enriching SNPs than do the *E. coli*.

However with the Ti/Tv ratio there is a little less certainty of the differences between the observed *Shigellae* values and the bootstrapped *E. coli* (table 3.6.6b). The differences being largely due to there being more of an overlap in the range of the extant *E. coli* and *Shigellae* ratios, most notably in the case of *S. sonnei* which is significantly greater than EcA, not significantly different to EcB, EcC and EcD and significantly less than EcE, given that EcA and EcE represent the lowest and highest *E. coli* Ti/Tv ratios respectively and *S. sonnei* shows a consistently *E. coli* like pattern of polymorphism profile results, it is to be expected that *S. sonnei* should show no difference from the 'mid-range' *E. coli*. Additionally *S. boydii*, based on percentile values, shows no difference from EcA, which is due to their similar Ti/Tv ratio (2.370 & 2.518 respectively). Similarly *S. dysenteriae* shows no difference from EcC, which again is due to similar Ti/Tv ratio (2.707 & 2.811 respectively).

Overall the bootstrap analysis confirms the patterns of relationship seen in the plots with respect to divergence time in Log(Number of SNPs) and the scatter plots from the residual analysis that the *Shigellae* and the *E. coli* are more separated by the +AT/+GC ratio than

by the Ti/Tv ratio, whilst at the same time both show a clear trend of a higher value in the *Shigellae*, excepting *S. sonnei* which shows a more *E. coli* like pattern.

	EcA	EcB	EcC	EcD	EcE
Sb	100	100	100	100	100
Sd	100	100	100	100	100
Sf	100	100	100	100	100
Ss	100	100	100	100	100

Table 3.6.6a – Observed Shigellae +AT/+GC ratios as a percentile of 1000 bootstrapped *E. coli* +AT/+GC Ratios, showing where the *Shigellae* are significantly higher (red) lower (green) or where there is no significant difference (blue).

	EcA	EcB	EcC	EcD	EcE
Sb	8.4	0	0	0	0
Sd	100	1	11.9	0	0
Sf	0	0	0	0	0
Ss	100	75.8	94.1	45.4	0.6

Table 3.6.6b – Observed Shigellae Ti/Tv ratios as a percentile of 1000 bootstrapped *E. coli* Ti/Tv Ratios, showing where the *Shigellae* are significantly higher (red) lower (green) or where there is no significant difference (blue).

3.7 – Summary of Results

- Time dependant purging of more deleterious SNPs is evident in both the d*N*/d*S* ratio and the proportion of SNPs occurring at fourfold degenerate (Q) sites.
- Differences between *E. coli* and *Shigellae* are consistent across functional categories.
- Principal component 1, representing 43% of the variation in the polymorphism profiles, shows a clear trend in score with respect to time.
- The two metric ratios recovered from PC1 both show time dependant purging of the more deleterious SNPs.
- Regression residual analysis of the metric ratios at non-fourfold (NQ) sites and the bootstrap comparison of metric ratios show separation of *E. coli* and *Shigellae*.
- *S. sonnei*, shows consistently more *E. coli* like patterns of polymorphism, highlighted by clustering with the *E. coli* rather than the other *Shigellae* in the regression residual plots.

3.8 – Discussion

It has been noted previously that the action of selection upon a mutation is not instantaneous, nor is it equal for all mutations. There is in fact a notable time delay between the generation of a mutation within a genome and it being purged or more rarely accepted, by selection. This has been characterised by the observation of a divergence time dependent trend in the d*N*/d*S* ratio (Rocha, Maynard Smith et al. 2006) highlighting that the more closely two genomes or sequences are related the greater the proportion of slightly deleterious (nonsynonymous) mutations that will be observed, as selection has yet to act upon them – thus the patterns will be more akin to the mutation bias. As more divergent sequences are compared the patterns observed will approach the selective bias as selection will have had a greater amount of time to act. The trajectory or slope of this time dependent trend is determined by the strength of purifying selection which itself is dependent upon both the effective population size of the organism (N_e) and selective coefficient of the mutation (s) (Ohta 1973; Rocha, Maynard Smith et al. 2006).

3.8.1 – Trends Observed in Ti/Tv and +AT/+GC

When comparing the metric ratios, along with bootstrapped confidence intervals, there is little observable difference between the taxa, especially in the ratio of transitions to transversions (figures 3.6.2 a & b). This reinforces the need to consider the time dependant trajectories of the metric ratios as opposed to simply their extant values.

The trend in the Ti/Tv ratio fits with the time dependant action of selection, given that no significant trend was observed with time at fourfold degenerate sites it is reasonable to relate this trend to effects on amino acid encoded, indeed twofold degenerate codons are linked at the third codon position by transitions, thus correspondingly a transversion at this site is nonsynonymous and likely to be selectively purged over time. In addition to this nonsynonymous transversions tend to result in less conservative, and hence more selectively costly, amino substitutions than nonsynonymous transitions (Freeland and Hurst 1998; Zhang 2000) as a result it is again the transversions that would be preferentially purged.

The selective purging of AT enriching polymorphisms is initially a little less intuitive than the purging of transversions, as AT richness does not have an obvious link to amino acid encoding. However it has been shown there is a correlation between the AT content of a sequence/genome and the metabolic cost of the amino acids encoded (Akashi and Gojobori 2002; Rocha and Danchin 2002). In addition to this direct effect, there are a number of features on which AT content has an effect which do not directly relate to protein coding, which is supported by the consistency of the trend in +AT/+GC between both non-fourfold and fourfold degenerate sites.

One potential cause of the trend towards purging of AT enriching polymorphisms is the maintenance of a particular codon bias, whilst a given synonymous nucleotide change may not alter the amino-acid encoded it may alter codon usage away from the optimal codons, incurring a selective cost in doing so. Examination of the codon usage tables for *E. coli* and *Shigella* (from the "Codon Usage Database" at http://www.kazusa.or.jp/codon/) reveals that in four-fold degenerate codons (used to avoid biases associated with examination of two, three and six-fold degenerate codons) the usage of codons with a G or C at the synonymous third codon position is greater than the usage of codons with an A or T at the same location, GC terminating codons representing between 52 and 71% of the codons used in a given synonymous subset (figures from *E. coli* O157:H7 sakai and are indicative of the trend across all the genomes in this dataset). This strongly suggests that the preservation of optimal codon usage is at least part of the impetus for the purging of +AT polymorphisms.

Additional features which may also play a contributory role to the deleterious nature of AT enrichment are the stability of AT pairings versus that of GC pairings (2 vs 3 hydrogen bonds for the former and latter respectively) which has implications for the overall stability of the base pairing in regions of extreme nucleotide bias. However stability is also of high concern with the structure of RNAs, the stability of the structure of molecules such as tRNA relies on the strength of bonding in complementary sections (Galtier and Lobry 1997). Again these effects would render an unfortunately located GC→AT mutation relatively costly and would favour its removal by purifying selection.

3.8.2 – Examination of Differences between *Shigellae & E. coli*

Detailed examination of the time dependant trend for all the analyses (d*N*/d*S*, PC1, Ti/Tv & +AT/+GC) reveals a contrast between the patterns observed in the *Shigellae* and the *E. coli*, with the former consistently showing greater proportions of more deleterious polymorphisms than the latter, even controlling for the relative effects of divergence time. These differences are clear in the XY scatter plots of the residuals to the regression lines with respect to time for both Ti/Tv and +AT/+GC at both NQ and Q sites (figures 3.6.5 a & b) as well as the bootstrap results for metric ratios, both of which show a clear separation of the *Shigellae* from the *E. coli*, the notable exceptions being the Ti/Tv residuals at Q sites which show no distinction, which is expected, and the consistent *E. coli* like patterns associated with *S. sonnei*. The consistency of the differences in d*N*/d*S* ratio across functional category of gene lends further credence to the differences overall, corresponding to greater d*N* in the *Shigellae*.

One potential explanation for these differences observed between the *E. coli* and the *Shigellae* is that the latter are under increased positive/diversifying selection owing either to host immune response (Wirth, Falush et al. 2006) or indeed due to attack by Phage or other Bacteria (Petersen, Bollback et al. 2007). Whilst this model would account for higher d*N*/d*S* ratios in the *Shigellae* it is highly unlikely that all 2098 genes analysed would be under diversifying selection simultaneously as they are neither all immunogenic nor all involved in defence against bacterial predators of phage. Specifically the comparison of d*N*/d*S* ratios between *E. coli* and *Shigellae* across various groups of genes by function excludes the possibility of adaptation to host immune response as the most likely immunogenic category of genes (Cell Envelope) shows no greater difference in d*N*/d*S* than does the vast majority of the other categories. It would also be expected that escape from attack by Phage or Bacterial predators would also require the adaptation of these genes, excluding these possibilities as well. It is worth noting however that whilst there is no evidence from this data of widespread positive selection, it has been reported that the loss of the function of the *cadA* gene (Lysine Decarboxylase) was the result of positive

selection, based upon the negative effects on pathogenicity of *S. flexneri* when this gene is reintroduced (Maurelli, Fernandez et al. 1998).

Hypermutability is also a potential reason for the surfeit of nonsynonymous polymorphisms seen in the *Shigellae* as a greater rate of mutation would result in the *Shigellae* showing a pattern of polymorphisms more akin to the mutation bias than the pattern seen in the *E. coli*. However work by Rocha, Cornet and Michel (Rocha, Cornet et al. 2005) has shown that the *Shigellae* possess a complete repertoire of homologous repair genes, as compared to those identified in the *E. coli*. Furthermore, a check of a list of 72 stress response genes identified (Rocha, Matic et al. 2002), which includes genes directly involved with DNA repair and genes implicated in hypermutation, revealed that all genes were present save for *ung* which, in *S. flexneri* 2a 301, contains a stop codon near the C-terminus. *ung* encodes uracil DNA glycosylase and is involved in the removal of Uracil residues resulting from the deamination of Cytosine and so is an important part of the DNA repair machinery. However, this mutation is absent from the *S. flexneri* strain 2a 2457T and so is either a sequencing error or is so recent that it will have had negligible influence on the patterns of polymorphisms observed, consequently hypermutators are unlikely to account for the patterns of polymorphism observed.

The nature of the intracellular environment the *Shigellae* inhabit could also impose strong mutation biases on the *Shigellae*, specifically in the case of the relatively higher proportion of AT enriching polymorphisms. The intracellular environment is relatively rich in both adenine and thymine, which may impose a strong usage bias (Rocha and Danchin 2002), in addition to this a mutational bias towards AT in enteric bacteria has been observed (Ochman 2003). As intriguing as these factors are as an explanation for the +AT/+GC differences between the *E. coli* and *Shigellae* they don't account for the decrease of the +AT/+GC ratio with time, additionally a mutational bias towards AT cannot intuitively explain the trends also observed in Ti/Tv and d*N*/d*S*. So whilst the mutation bias towards AT is likely a contributing factor to the ongoing genomic changes in the *Shigellae* it is not an explanation of the observed evolutionary differences to the *E. coli*.

3.8.3 – Reduced Purifying Selection as the 'Best Fit'

The only model which can account for all of the observed results is that of reduced purifying selection in the *Shigellae*. However reduced purifying selection encompasses two different processes which are somewhat troublesome to tease apart: Increased genetic drift, a result of inefficient selection due to low effective population (N_e) size, in other words an increased chance that any given mutation will reach fixation by stochastic processes before it is processed by selection, or relaxed selection due to more constant and favourable environmental conditions, such that the selective cost (*s*) of any given mutation is reduced. In reality the ecological niche adopted by the *Shigellae* is likely to have imposed both a reduced effective population size, due to the physical constraints of intracellular replication and potentially through selective sweeps upon acquisition of the pINV plasmid or other genetic traits beneficial/adaptive for intracellular life, as well as lowered the selective cost of mutations due to the relatively stable and predator free nature of the intracellular environment of the host cell.

Even given that it has been argued that relaxed selection is unlikely to affect all loci simultaneously and/or equally (Wernegreen and Moran 1999), this data provides no conclusive evidence in favour of either relaxed selection or increased genetic drift, in all likelihood both mechanisms have a contributory role, with the balance between the two processes varying between loci or within a given locus.

There is one category of genes which shows a markedly higher rate of nucleotide evolution; fat metabolism. There are two conflicting potential explanations for the strong differences observed both share a common cause; that of the adoption of an intracellular pathogenic niche resulting in access to the abundant energy stores of the host cell. Potentially purifying selection may be acting on these genes as maintaining an unnecessary metabolic pathway is selectively costly, at the same time however the fact that the pathway has become redundant will reduce the selective cost of mutations in genes of that pathway, therefore reduced purifying selection is also a potential explanation. The *Shigellae* show a surfeit of both AT enriching SNPs (mean of *E. coli* minus mean of *Shigellae* equals -0.660) and Transversions (mean of *E. coli* minus mean

of *Shigellae* equals 0.351) which is more in keeping with reduced purifying selection, than positive or diversifying selection. The latter would only be of benefit if the genes are inactivated, rather than the production of partially or non-functional proteins through the introduction of relatively non-conservative and potentially metabolically costly amino-acid changes represented by the increased proportion of Transversions and relative AT enrichment.

3.8.4 – *S. sonnei* as a Special Case

In general metrics and analyses which show a clear time dependence highlight the patterns of evolutionary change observed in *S. sonnei* as more reminiscent *E. coli* than the other *Shigellae*. It is important to note that this is not an effect relating to the phylogenetic relationship of the taxa used, as *S. sonnei* falls within a cluster with *S. boydii* and *S. flexneri*. A likely explanation for this distinction is the recent adoption of the intracellular pathogen niche by *S. sonnei*. It has been estimated that *S. sonnei* acquired its pINV plasmid around 10,000 years ago (Shepherd, Wang et al. 2000), far more recently than any of the other *Shigellae*, in agreement with this Karaolis et al (1994) noted that *S. sonnei* showed very little sequence diversity further supporting the recent origin.

The patterns of genome degradation, associated with reduced purifying selection, seen in *S. sonnei* suggest that it represents a 'halfway house' between *E. coli* and full-blown *Shigellae* (Hershberg, Tang et al. 2007) specifically in the numbers of genes lost in comparison to *E. coli* K-12 – 255 for *S. sonnei* versus an average of 151 (132-180) and 396 (371-543) for *E. coli* and the other *Shigellae* respectively. This is somewhat akin to the status afforded to Enteroinvasive *E. coli* (EIEC), with *S. sonnei* being the only *Shigella* to fall into a clade with only EIEC strains (Lan, Alles et al. 2004).

S. sonnei also has a slightly different epidemiology; it is responsible for roughly 78% of sporadic outbreaks of shigellosis in industrialised countries with high levels of hygiene (Pal 1984; Sultana, Mizanur et al. 2002). This in turn would suggest that it has an environmental host, in fact *S. sonnei* (along with *S. dysenteriae*) has been observed in

Acanthomoeba, specifically *A. castelanii* (Jeong, Jang et al. 2007; Saeed, Abd et al. 2008).

Overall this suggests that *S. sonnei* has had less time to accumulate the level of deleterious polymorphisms observed in the other *Shigellae* and has the potential to maintain a larger effective population size though survival in a non-human host, which in turn provides opportunities for potentially restorative horizontal gene transfer with *S. dysenteriae*, a benefit which would be largely one-way given the extreme clonality of the *S. sonnei* population.

Chapter 4 – Patterns of Amino-Acid and Codon Position Nucleotide Change in the *Shigellae*

4.1 – Introduction

4.1.1 – Consideration of Codon positions

Given that each of the codon positions has differing levels of degeneracy, with the third codon position being the most degenerate, followed by the partially degenerate first position and the second position being nondegenerate, they will be under different levels of selective constraint. Investigation of the patterns of nucleotide change at each codon position is therefore a logical extension of the nucleotide patterns examined in Chapter 3.

These differences in selective constraint or degeneracy, can explain the patterns of GC content of each codon position versus genomic GC content observed by Muto and Osawa (1987). The figure below (4.1.1a) shows how the GC content of each codon position varies with overall genome GC content. The third codon position, being the most degenerate and therefore under the least selective constraint, shows the greatest variation in GC content, broadly in line with the GC content of the whole genome. The partially degenerate first position shows some variation in GC content, correlated to the whole genome GC content and the second codon position, being absolutely nondegenerate and therefore under the highest selective constraint, shows minimal variation of GC content with whole genome GC content.

This in turn raises the likelihood that there may also be observable differences in the amino acid polymorphisms of the shared proteomes of *E. coli* and *Shigellae*. In light of observations by Jordan et al (2005) and Hurst et al (2006) it is reasonable to assume that the same time dependence of the purging of more deleterious changes evident in the nucleotide polymorphism profiles will be observable, as a greater proportion of more deleterious changes at the initial stages of divergence, in the amino acid polymorphisms.

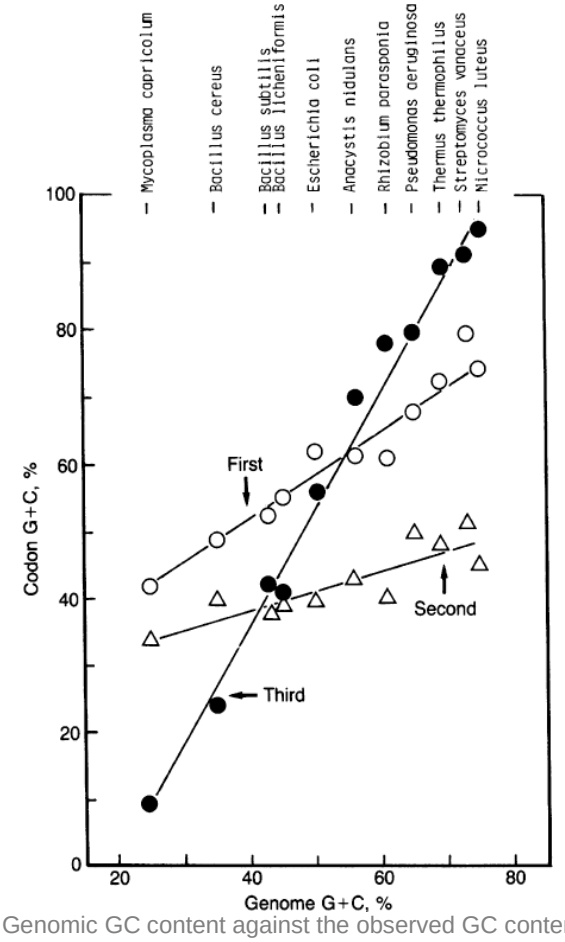


Figure 4.1.1a – Plots of the Genomic GC content against the observed GC content at each codon position for various prokaryotic taxa. Adapted from Muto & Osawa (1987).

4.1.2 – Estimation of Metabolic Costs of Amino Acids

The estimation of the metabolic costs for each amino acid was performed by Akashi and Gojobori (2002), using a method which is in turn an adaptation of that used by Craig and Webber (1998). Under this approach the metabolic cost of each amino acid is based upon the steps involved in its synthesis via three pathways (Embden-Meyehof-Parnas, Tricarboxylic acid and Pentose Phospate) as well as the cost of any precursor molecules, which are in turn averages of growth on three different carbon sources – Glucose, Acetate and Malate. The calculation was made based upon *E. coli*, assuming that the availability of nitrogen (in the form of NH_4) and sulphates was non-limiting; the cost of reduction of the sulphates to H_2S for Met and Cys was included. The resulting values are in high energy phosphate bond (~P) equivalents, assuming that a single available hydrogen atom (in the form of NADH, NADPH or $FADH_2$) is equivalent to 2 ~P and are shown in table 4.1.2a.

Amino Acid		~P in ATP and GTP	Available H atoms in NADH NADPH & $FADH_2$	Total Cost (~P Equivalent)
Single Letter Code	Short Code			
A	**Ala**	1	5.3	11.7
C	**Cys**	7.3	8.7	24.7
D	**Asp**	1.3	5.7	12.7
E	**Glu**	2.7	6.3	15.3
F	**Phe**	13.3	19.3	52
G	**Gly**	2.3	4.7	11.7
H	**His**	20.3	9	38.3
I	**Ile**	4.3	14	32.3
K	**Lys**	4.3	13	30.3
L	**Leu**	2.7	12.3	27.3
M	**Met**	9.7	12.3	34.3
N	**Asn**	3.3	5.7	14.7
P	**Pro**	3.7	8.3	20.3
Q	**Gln**	3.7	6.3	16.3
R	**Arg**	10.7	8.3	27.3
S	**Ser**	2.3	4.7	11.7
T	**Thr**	3.3	7.7	18.7
V	**Val**	2	10.7	23.3
W	**Trp**	27.7	23.3	74.3
Y	**Tyr**	13.3	18.3	50

Table 4.1.2a – A list of the amino acids and their corresponding metabolic costs, including the breakdown by High Energy Phosphate Bond (~P) and available Hydrogen atom (H), according to (Akashi and Gojobori 2002)

4.1.3 – Aims & Conclusions

Here I aim to further the results observed in Chapter 3 by examining the differences in the derived metric ratios between the *Shigellae* and *E. coli* at each codon position as well as analysing the amino acid polymorphisms evident between the core genomes and the associated metabolic cost. I observe a pattern of nucleotide differences more complex than in Chapter 3 but which still show trends indicative of the time dependant purging of more deleterious changes, and indicate reduced purifying selection in *Shigellae*. I also observe patterns associated with the time dependant purging of more deleterious changes in the amino acid polymorphisms.

4.2 – Codon Position Distribution of SNPs

4.2.1 – Initial Analysis

Comparison of the internal branch estimation methods (table 4.2.1a) reveals that the level of agreement is broadly similar to that observed earlier. The two methods correlate strongly at 1st and 3rd codon positions, but not at the second codon position which is borderline non-significant (p = 0.055). It is possible that that relative paucity of SNPs at the second codon position is a key to the discrepancy between the two methods, there simply may not be sufficient information to reliably estimate the total number of SNPs present.

Method	Internal Branch	Number of SNPs			
		1st Codon Position	2nd Codon Position	3rd Codon Position	All Sites
PAML	iA	261	140	1404	1805
	iB	147	53	1128	1328
	iC	297	93	2491	2881
	iD	250	83	2253	2586
	iE	712	219	7317	8248
TBS	iA	262	139	1419	1819
	iB	329	140	2104	2573
	iC	358	121	3003	3481
	iD	259	85	2301	2645
	iE	691	330	9536	10826

Table 4.2.1a – The total number of SNPs at each nucleotide site type for each internal branch under both methods. The calculated values are the mean of all possible ways of estimating the branch values.

The general trend evident from the total number of SNPs at each type is consistent with the known degeneracy of the nucleotide sites, that is the vast majority of SNPs are at the highly degenerate third codon position, followed by the less degenerate but generally conservative (in terms of encoded amino-acid) first codon position and lastly the absolutely nondegenerate second position (table 4.2.1b). This is unsurprising given the exclusively nonsynonymous nature of changes the second codon position, as well as the non-conservative amino-acid changes that result from second position SNPs, the extreme examples being N→A mutations in codons beginning in T and ending G or A, which results in conversion to a STOP codon. Selective biases against the more deleterious polymorphisms at first and second sites would result in the enrichment for third codon position polymorphisms observed.

Species / Strain Code	Number of SNPs			
	1st Codon Position	2nd Codon Position	3rd Codon Position	All Sites
EcA	673	357	3520	4550
EcB	857	455	4206	5518
EcC	868	429	4302	5599
EcD	1507	555	11134	13196
EcE	710	357	4013	5080
Sb	905	717	2563	4185
Sd	1823	1267	5598	8688
Sf	1283	959	3586	8528
Ss	718	482	2415	3615
iA	262	139	1419	1819 (4810)
iB	329	140	2104	2573 (7042)
iC	358	121	3003	3481 (9125)
iD	259	85	2301	2645 (8426)
iE	691	330	9536	10826 (15906)

Table 4.2.1b – The total number of SNPs at each codon position within the sequences. All site values for internal branches, bracketed in grey, represent the divergence "Time" of the SNPs – they are a reflection of the phylogenetic depth of the internal branches (calculated as in methods – 2.8.3).

4.2.2 – Codon Position Bias over Time

The clear trend with time is towards the purging of the more deleterious second and first codon position SNPs, resulting in them accounting for a lower proportion of the total number of SNPs. As expected given the relative degeneracy of each position and their relative selective costs, there is a significant trend of enrichment of SNPs at third codon positions with time; consequently there are significant trends for the purging of both first and second codon position SNPs with time. These effects are largely a reflection of the purging of more deleterious SNPs, typified by the time dependence of the d*N*/d*S* ratio observed in Chapter 3 (3.4.1).

In line with earlier observations in Chapter 3, the *Shigellae* show no significant trend towards enrichment for third codon position SNPs; Adj-$R^2 \approx 0$ and $p > 0.10$ for the trend based upon the *Shigellae* alone. In addition, as can be seen in figure 4.2.2b, there is no evidence for the relative enrichment of 1st position SNPs versus 2nd position SNPs with time in the *Shigellae*, whilst the *E. coli* show a strong trend, indicating the preferential purging of 2nd position SNPs.

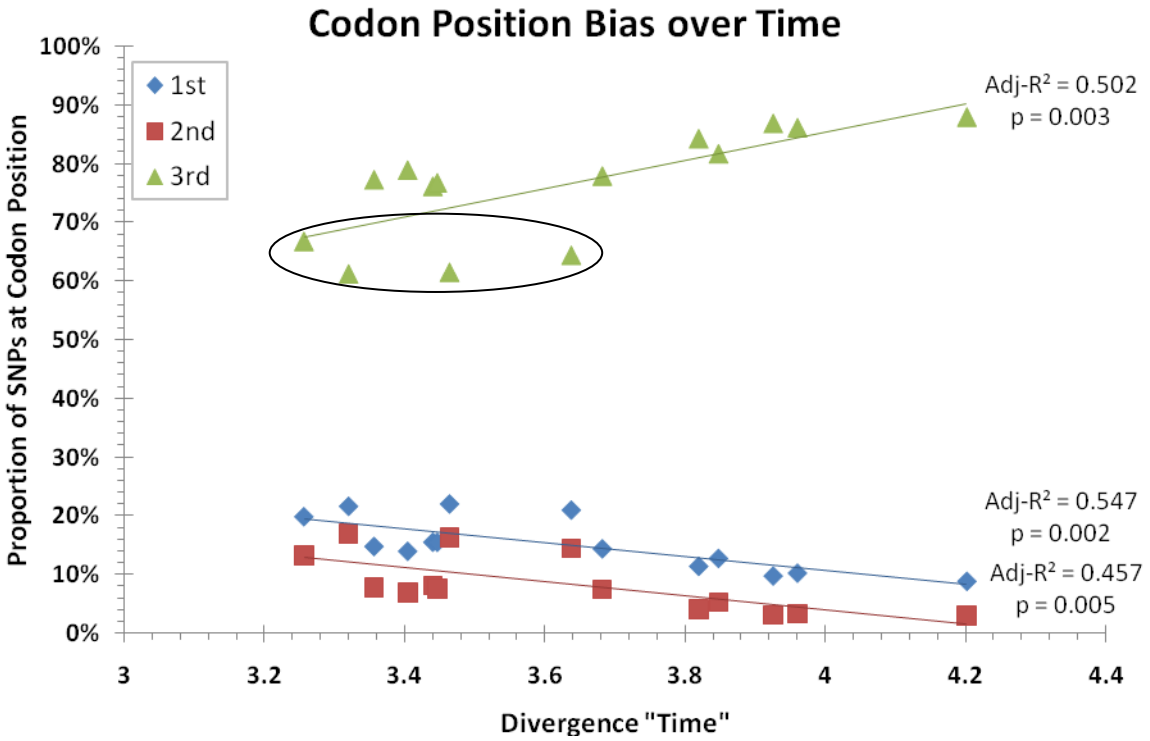


Figure 4.2.2a – The proportion of SNPs at any given codon position against divergence "time" for each taxon, showing linear least-squares regression lines with Adjusted R^2 and p values. The ringed points (for 3rd site positions) are the *Shigellae*.

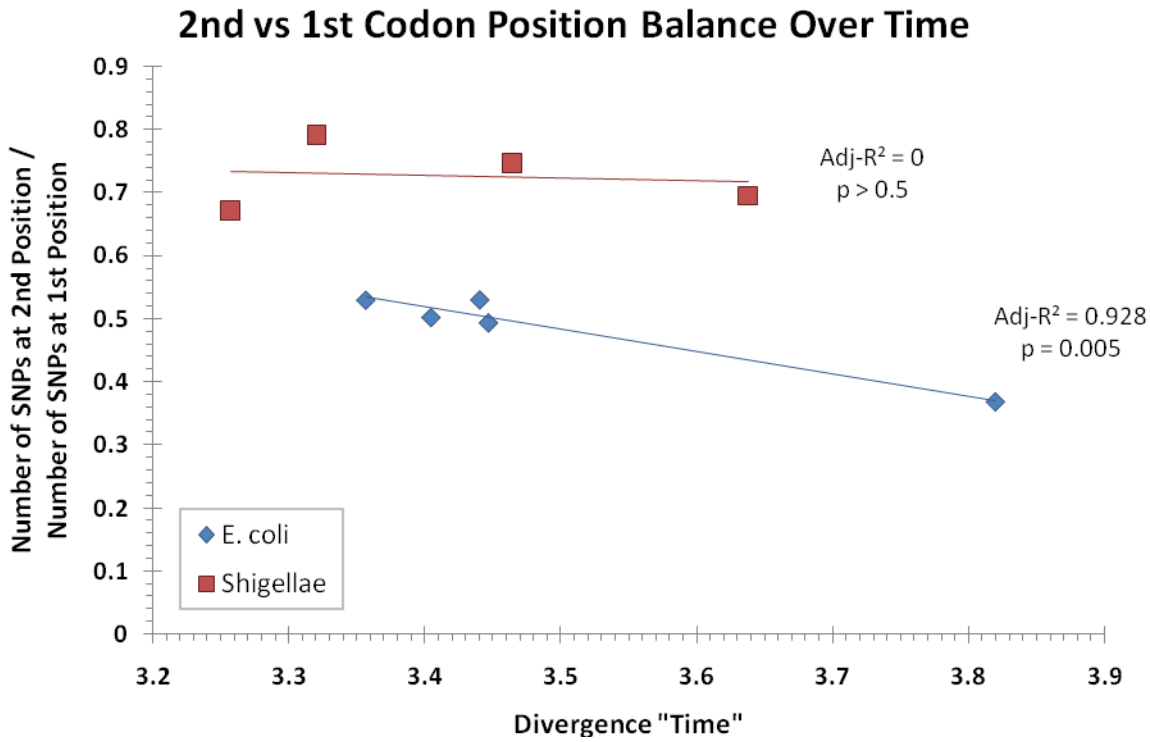


Figure 4.2.2b – The ratio of the number of SNPs at the 2nd position to the number of SNPs at the 1st position with time, showing linear least squares regression lines with adjusted R^2 and p values.

4.3 – Metric Ratios each Codon Position

4.3.1 – Metric Ratio Values

In line with the principal components analysis in Chapter 3, the metric ratios of +AT/+GC and Ti/Tv were chosen and calculated using the normalised polymorphism profiles for each of the 9 taxa and 5 internal branches, at each codon position, 'All Site' values are shown for reference. The ratios are listed below (Table 4.3.1a & b for +AT/+GC & Ti/Tv respectively).

Taxon / Branch ID	Divergence Time	Ratio of AT enriching to GC enriching SNPs (+AT / +GC)			
		All Sites	1st Codon Position	2nd Codon Position	3rd Codon Position
EcA	3.357	2.320	1.967	3.462	1.871
EcB	3.441	2.348	2.406	2.578	1.911
EcC	3.447	2.179	2.070	2.830	1.784
EcD	3.819	1.708	1.506	2.683	1.373
EcE	3.405	2.412	2.215	2.593	1.996
Sb	3.321	2.878	2.893	3.149	2.431
Sd	3.638	2.485	2.710	2.886	2.001
Sf	3.464	2.871	2.608	3.051	2.513
Ss	3.257	2.677	2.932	3.455	2.075
iA	3.682	2.326	2.036	1.790	2.051
iB	3.847	2.064	1.886	2.992	1.653
iC	3.960	1.759	1.557	2.207	1.441
iD	3.926	1.847	1.717	1.857	1.498
iE	4.201	1.235	0.981	1.659	1.013

Table 4.3.1a – The +AT/+GC ratio for each Taxon and internal branch (estimated using TBS) for All sites and each codon position.

Taxon / Branch ID	Divergence Time	Transition/Transversion Ratio (Ti / Tv)			
		All Sites	1st Codon Position	2nd Codon Position	3rd Codon Position
EcA	3.357	2.518	2.117	1.633	2.671
EcB	3.441	2.888	2.717	1.791	3.090
EcC	3.447	2.811	2.551	1.592	2.971
EcD	3.819	2.978	3.462	1.795	2.954
EcE	3.405	3.291	3.421	2.132	3.387
Sb	3.321	2.370	1.999	1.907	2.650
Sd	3.638	2.707	2.485	2.265	2.833
Sf	3.464	2.276	2.000	1.564	2.598
Ss	3.257	2.949	2.610	2.411	3.116
iA	3.682	2.369	1.523	0.990	2.823
iB	3.847	2.773	3.620	1.525	2.734
iC	3.960	3.103	3.760	1.486	3.045
iD	3.926	2.980	3.416	1.256	2.984
iE	4.201	3.215	4.215	1.515	3.213

Table 4.3.1b – The Ti/Tv ratio for each Taxon and internal branch (estimated using TBS) for All sites and each codon position.

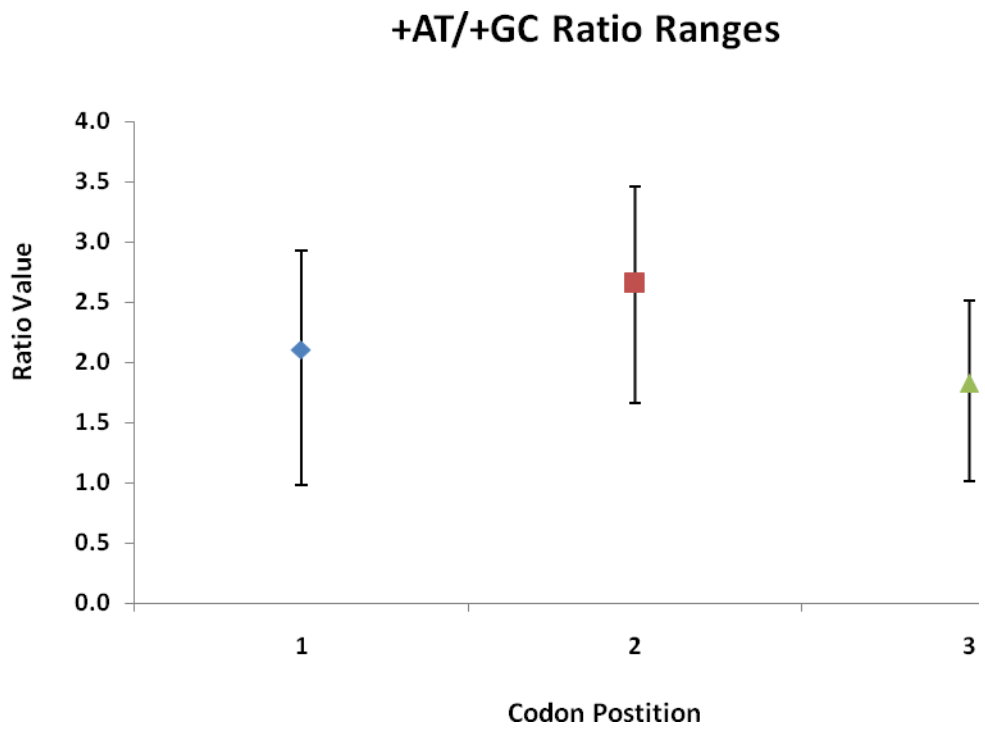


Figure 4.3.1a – The +AT/+GC ratio mean and range at each codon position based upon extant taxa and internal branches.

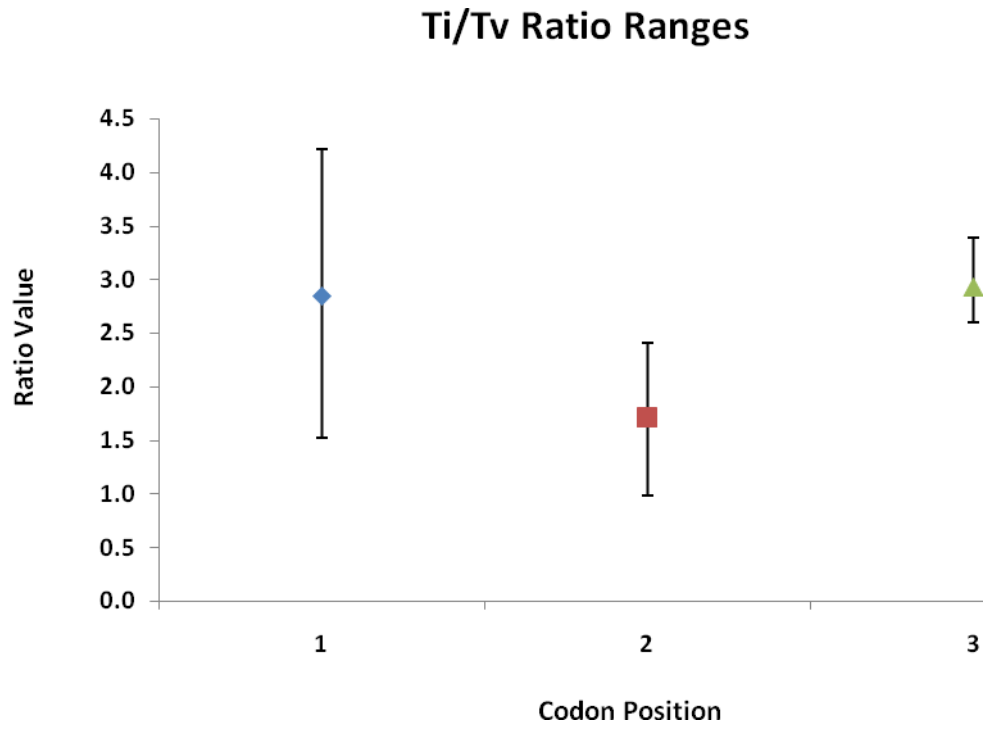


Figure 4.3.1b – The Ti/Tv ratio mean and range at each codon position based upon extant taxa and internal branches.

The +AT/+GC ratio is greater than one across all site types with the singular exception of iE at the first codon position, the extreme value likely reflecting the phylogenetic depth of iE (figure 4.3.1a). This surfeit of AT enrichment can in part be attributed to the relatively common C→T mutation.

Whilst the first and third codon positions show significantly different values in a paired t-test ($p = 0.002$), they are strongly correlated (Pearson's $r = 0.903$, $p = 0.000$), which suggests a possible common process responsible for their +AT/+GC ratios. Indeed examination of the AT content of both first and third sites reveals them to be AT poor (an average of 39% and 42% AT at first and third sites respectively), this results in a mutational bias towards AT enrichment given that there are more 'opportunities' for GC→AT than the reverse. In fact the more AT poor first codon position has a higher proportion of AT enriching polymorphisms than the less AT poor third codon position (mean +AT/+GC ratios of 2.106 and 1.829 for first and third sites respectively). It is unclear why the second codon position shows such strongly +AT biased patterns of polymorphism, especially given that it is AT rich (approximately 59% AT).

SNP Type		Codon Position		2^{nd} / 3^{rd}
		2^{nd}	3^{rd}	
+AT	C→A	7.1%	3.9%	1.82
	C→T	30.1%	29.6%	1.02
	G→A	20.8%	20.4%	1.02
	G→T	4.8%	4.8%	1.00
+GC	A→C	4.9%	2.5%	1.98
	A→G	8.7%	11.8%	0.74
	T→C	5.5%	12.5%	0.44
	T→G	2.2%	3.2%	0.69

Table 4.3.1c – A breakdown of the +AT/+GC ratio at 2^{nd} and 3^{rd} codon positions

A comparison of the polymorphism profiles of codon positions 2 and 3 (table 4.3.1c) shows a marked increase in the proportions of C→A & A→C at 2^{nd} positions, which essentially cancel each other out in terms of AT/GC balance. Additionally there is a notable decrease in the proportion of T→C, T→G and A→G, it is possible that constraints on amino acid properties are responsible for these observed differences in the pattern of changes at 2^{nd} positions – there is a known correlation between 2^{nd} position AT content

and hydrophobicity of the encoded amino acid; specifically the most hydrophobic amino acids having a T at the second position in their codons (Haig and Hurst 1991), the maintenance of hydrophobic amino acid residues at frequencies above their mutational equilibrium would result in both an AT rich sequence and a bias in the SNP profiles towards relative AT enrichment. Whilst the differences between 2nd and 3rd position do not show an increase in G/C→T SNPs they do show a marked decrease in T→G/C, reflecting a decreased purging of Ts.

The Ti/Tv ratio shows patterns consistent with the known inherent bias towards transitions – with the singular exception of iA at codon position 2, all sites show a ratio consistently higher than 1 (figure 4.3.1b). The Ti/Tv ratio values at the 1st and 3rd positions are not significantly different ($p = 0.608$) and are correlated (Pearson's $r = 0.610$, $p = 0.021$), strongly suggesting that similar or related processes are determining the balance of transitions and transversions at both sites.

In spite of the high degeneracy of the 3rd codon position, changes at these sites are not absolutely synonymous; amino acids which are encoded with twofold degeneracy have codons which are interlinked exclusively by transitions at the third codon position. Consequently a bias towards the purging of transversions would be expected as they are relatively more nonsynonymous at these sites.

At the 1st codon position the vast majority of changes are nonsynonymous, however amino acids with similar properties in terms of size or chemistry are more likely to share codons linked by transitions than transversions at the 1st position, as a result purifying selection would preferentially purge the more deleterious transversions at this position.

Second codon position changes are all nonsynonymous and most involve a change in the size or chemical properties of the amino-acid, such that there is little difference in selective cost between transitions and transversions and as a consequence the Ti/Tv ratio reflects the reported mutational bias in favour of transitions (Collins and Jukes 1994), the mean ratio of all 2nd site polymorphisms is 1.704.

4.3.2 – AT versus GC enrichment over time

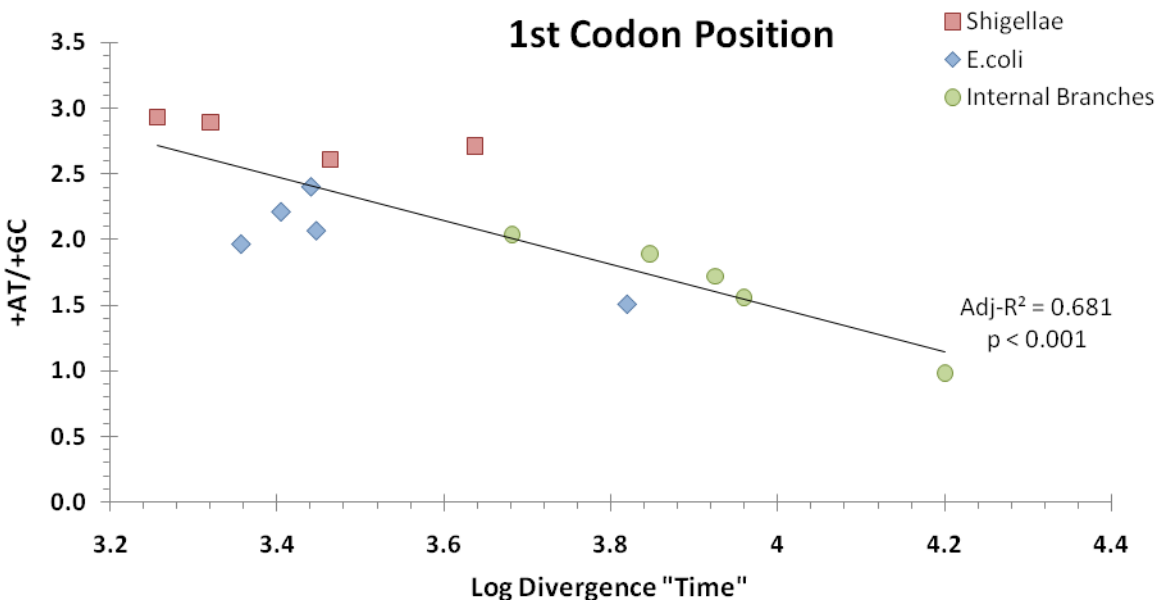
All three codon positions show strong significant downward trends with increasing divergence time (as indicated by their high adjusted R^2 values and low p values, see figure 4.3.2a,b,c&d), reflecting a gradual purging of AT enriching SNPs.

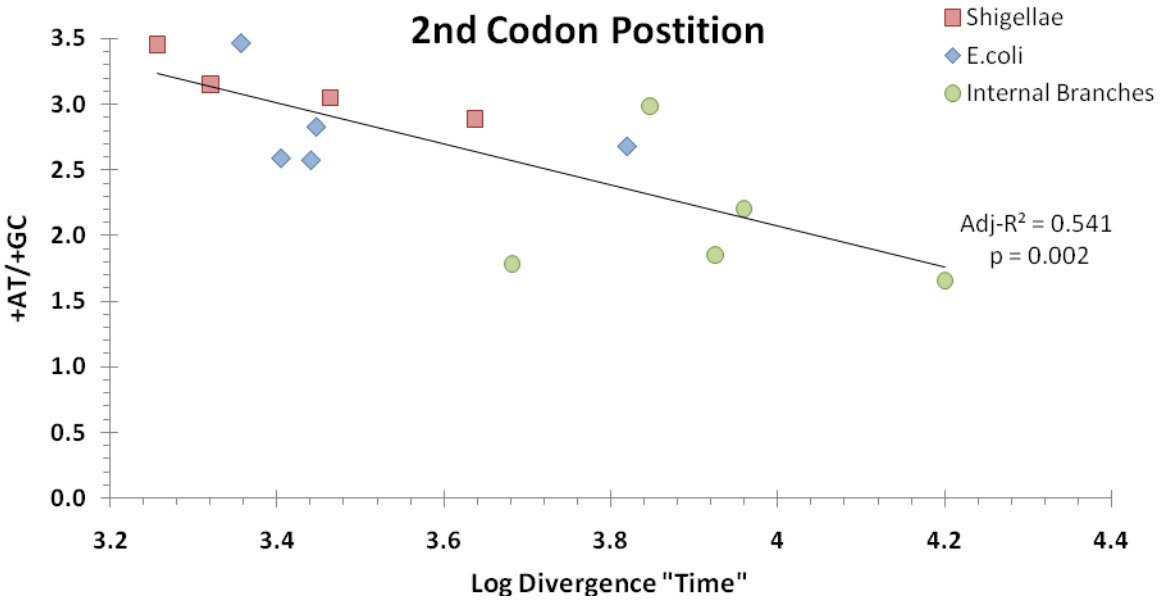
This holds with the general principal that AT enrichment is mildly deleterious, as it is known to have several potentially deleterious effects; including increasing the mean metabolic cost of proteins, a weak but evident trend observed by Heizer et al (2006), potentially reducing mRNA stability and altering mRNA secondary structure (Chamary and Hurst 2005) and can alter the codon bias of the sequence away from that preferred by the organism, introducing translational inefficiencies.

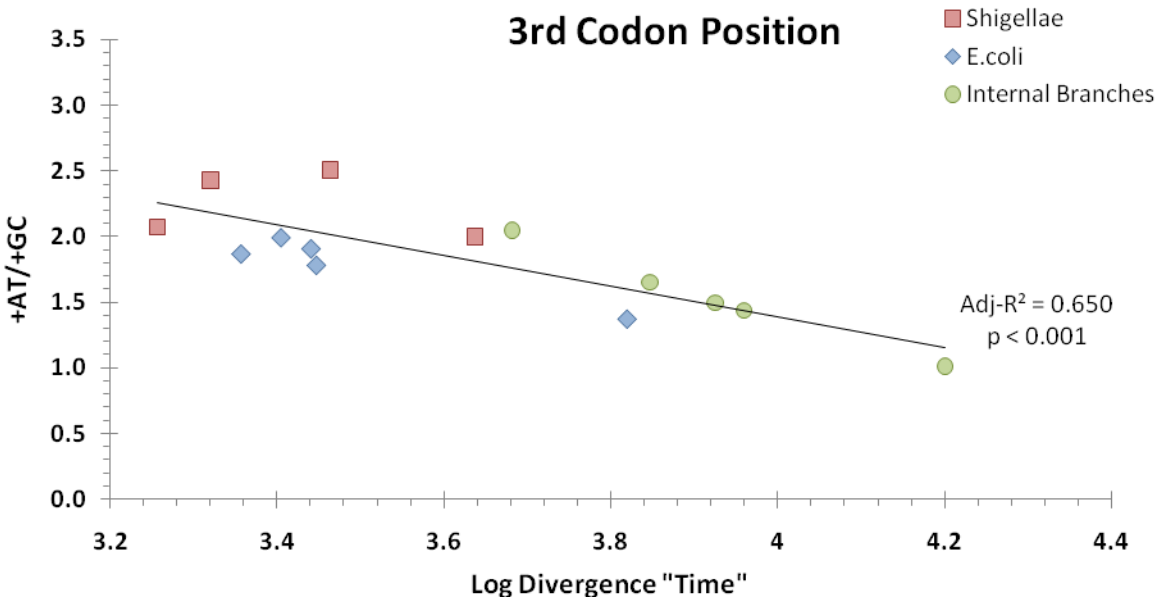
Both the first and third codon positions show effects which separate the *E. coli* from the *Shigellae* suggesting that there are different evolutionary trends or patterns in each group, this is reinforced by the mean gradients of the trendlines for each group; -1.35 for *E. coli* only, -0.53 for *Shigellae* only and -1.42 for all taxa including internal branches.

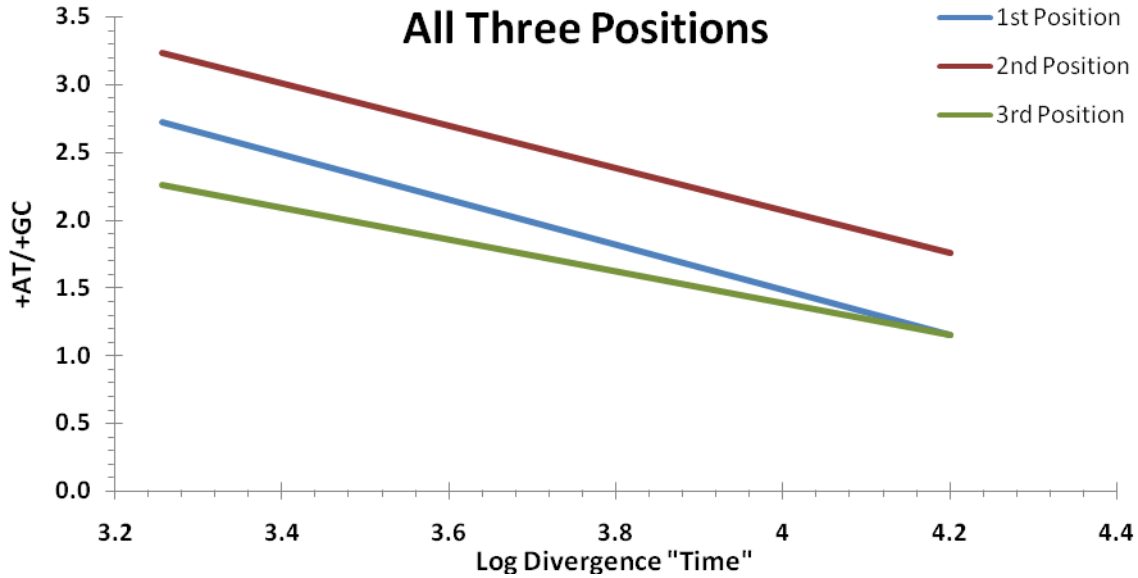
The second codon position however shows no clear separation of the *Shigellae* and *E.coli*, with the latter being relatively evenly distributed either side of the regression line and the former lying either on the regression line or near it. This is somewhat suprising giving the consistency of the differences observed at both fourfold degenerate and nonfourfold degenerate sites (Chapter 3).

There are many potential explanations for the apparent lack of difference between *Shigellae* and *E.coli* in the time dependant trend of +AT/+GC at the 2nd position; greater purifying selection on the second codon position of the *Shigellae*, or reduced purifying selection at the second position in the *E. coli*, or even that selective constraint at the second position is high enough for there to be minimal observable effect. However, it is also possible that the lack of differentiation observed is a consequence of the low number of SNPs at the second position (mean = 619) as compared to the first or third positions (means of 1038 and 4593 respectively), resulting in decreased power to resolve any differences which may be present.


1st Codon Position
Shigellae
E.coli
Internal Branches
+AT/+GC
Adj-R² = 0.681
p < 0.001
Log Divergence "Time"
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4


2nd Codon Postition
Shigellae
E.coli
Internal Branches
+AT/+GC
Adj-R² = 0.541
p = 0.002
Log Divergence "Time"
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4


3rd Codon Position
Shigellae
E.coli
Internal Branches
+AT/+GC
Adj-R² = 0.650
p < 0.001
Log Divergence "Time"
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4



Figures 4.3.2 a, b, c & d – The +AT/+GC ratio versus time at codon positions 1, 2 and 3 respectively. With **d** showing the three regression lines overlaid.

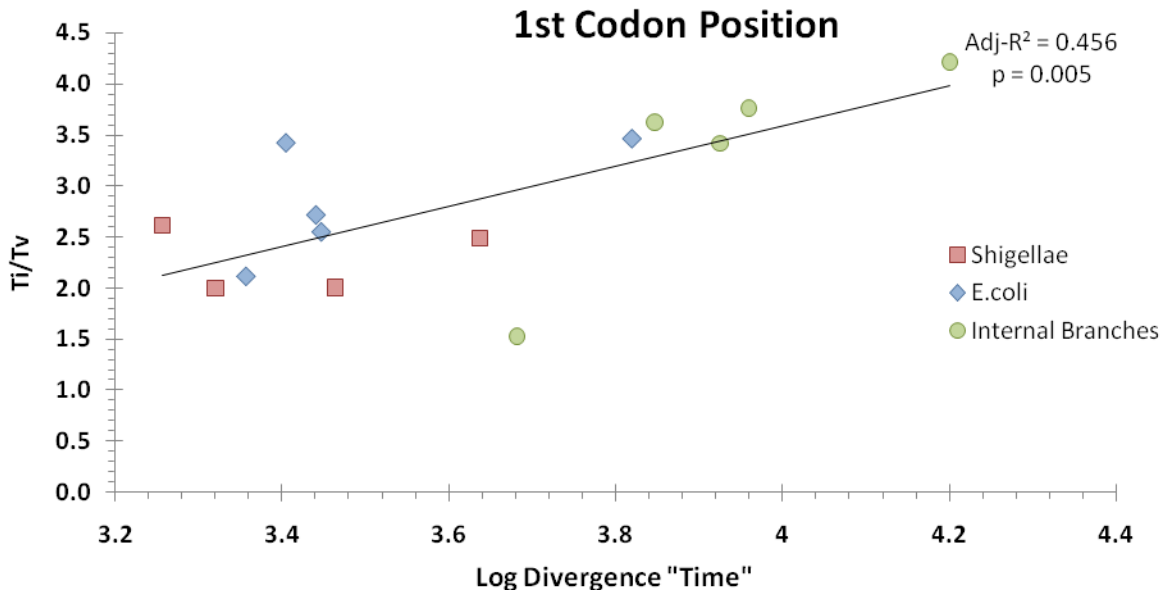
4.3.3 – Transitions versus Transversions over time

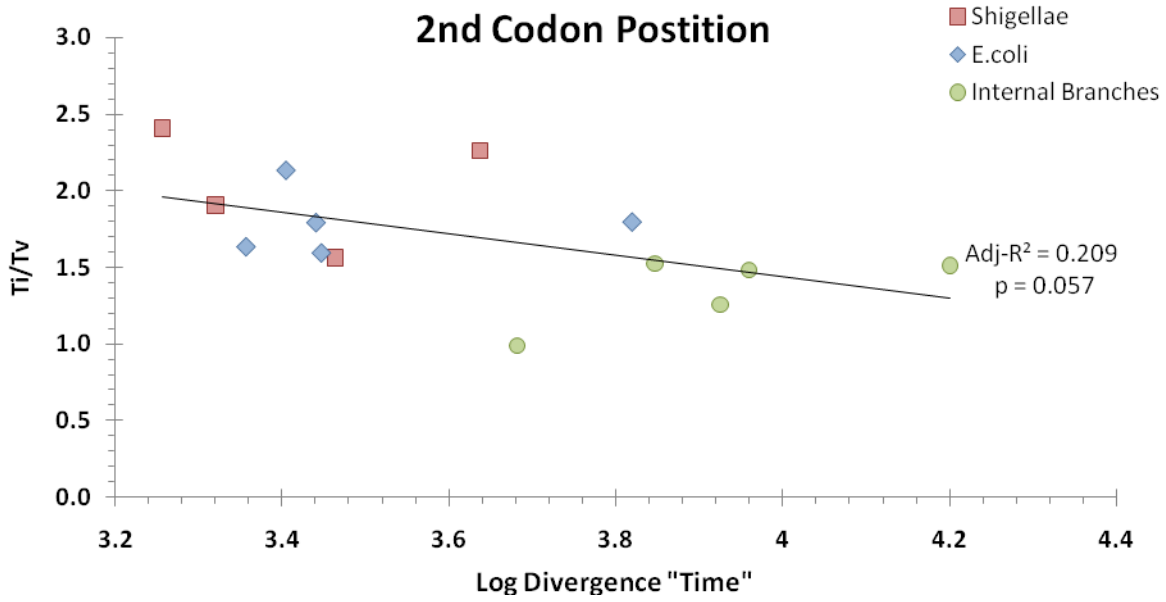
There is no consistent time dependant trend across all three codon positions. These differences likely reflect the relative levels of degeneracy of each codon position and the differential selective costs associated with transitions and transversions, the former resulting in more synonymous or conservative changes than the latter.

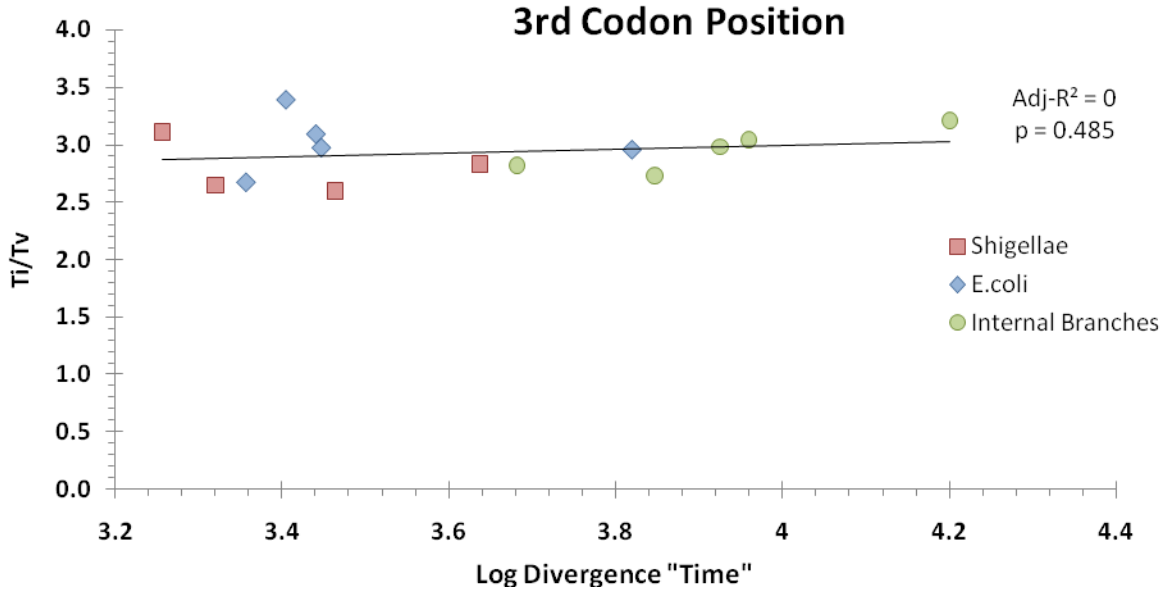
The first codon position shows a significant and strong trend towards an increasing proportion of Transitions with time; given that nonsynonymous Transitions result in more conservative amino acid changes than do nonsynonymous Transversions and that the majority of changes at first codon positions are nonsynonymous, it is likely that this reflects the purging of the more deleterious changes. Also evident is the separation of the *E. coli* and *Shigellae*; the former showing a trend not dissimilar to the overall trend shown below (figure 4.3.3a) – gradients of 1.96 and 1.98 for the *E. coli* and the dataset as a whole respectively. The *Shigellae* show no such trend with a regression gradient of 0.08, showing a minimal amount of change with time, likely reflecting the different lifestyle adopted by the *Shigellae*. This pattern mirrors that seen earlier at the NQ sites in Chapter 3 (specifically in figure 3.6.4b), reflecting the largely non-degenerate nature of both site classes.

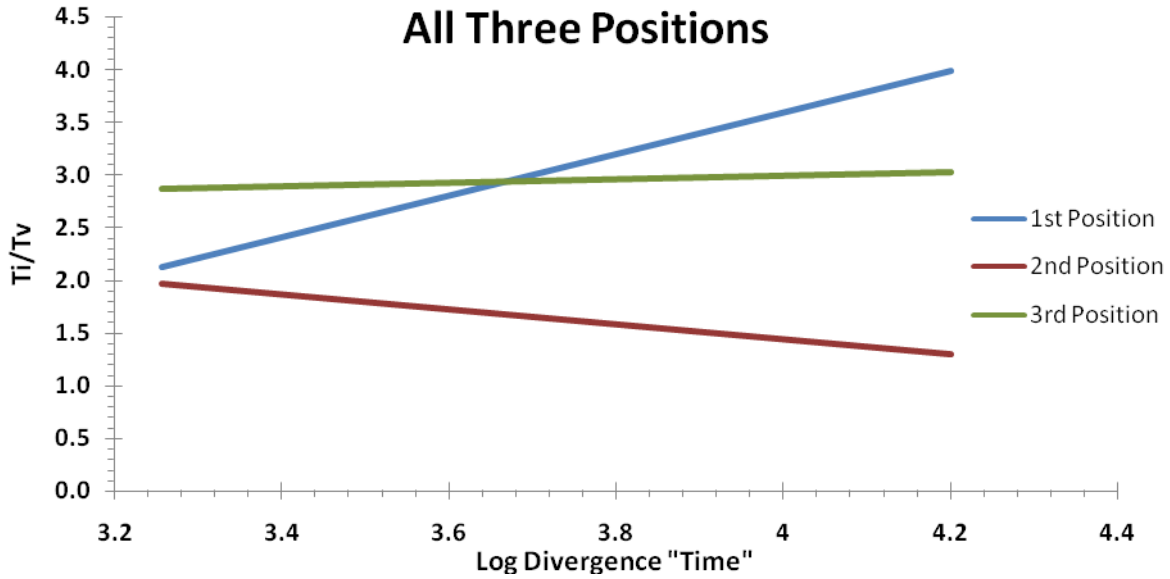
Whilst there is an apparent, albeit slight, trend at the second codon position, it is neither strong nor has strong statistical support, it is not surprising therefore that there is no obvious separation of the *E. coli* and *Shigellae*. This is possibly due to the absolute nondegeneracy of the second positions, which combined with the generally non-conservative nature of changes at this site, renders the vast majority of changes selectively equal and so whilst changes at this position will be preferentially purged as compared to other positions, most changes within this position will be treated equally. Additionally, as mentioned previously (4.3.2), there are relatively few SNPs observable at second position, limiting the statistical power of the analysis at this site and also potentially reducing the ability to resolve any differences between the *Shigellae* and *E. coli*.

Whilst the 3rd position is highly degenerate, it is not absolutely so and twofold degenerate sites at the 3rd position are synonymously connected exclusively by transitions, as such it would be expected that there would be a strong time-dependant selective bias against transversions at this position. Whilst there is a strong bias in the ratio towards transitions there is no observable time-dependence of this bias.


1st Codon Position
Adj-R² = 0.456
p = 0.005
Shigellae
E.coli
Internal Branches
Ti/Tv
Log Divergence "Time"
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4


2nd Codon Postition
Shigellae
E.coli
Internal Branches
Adj-R² = 0.209
p = 0.057
Ti/Tv
Log Divergence "Time"
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4


3rd Codon Position
Adj-R² = 0
p = 0.485
Shigellae
E.coli
Internal Branches
Ti/Tv
Log Divergence "Time"
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.2
3.4
3.6
3.8
4
4.2
4.4



Figures 4.3.3 a, b, c & d – The Ti/Tv ratio versus time at codon positions 1, 2 and 3 respectively. With **d** showing the three regression lines overlaid.

4.3.4 – Metric Ratio Summary Table

	+AT/+GC			Ti/Tv		
	1	**2**	**3**	**1**	**2**	**3**
Mean	2.106	2.656	1.829	2.850	1.704	2.933
Adj-R2	0.681	0.541	0.650	0.456	0.207	0
p	< 0.001	0.002	< 0.001	0.005	0.057	0.485
T-test *E. coli* vs *Shigellae*	0.002	0.232	0.071	0.154	0.333	0.224
(Including Ss within the *E.coli*)	(0.039)	(0.628)	(0.084)	(0.052)	(0.940)	(0.030)

Table 4.3.4a – Summary of the trends with respect to time at each codon position. The T-test values are p-values for a 2-Sample T-test of the residuals to the regression lines

It is clear from the table above that the differences between *E. coli* and *Shigellae*, especially Ti/Tv only become statistically significant when *S. sonnei* is included within the *E.coli* for the analysis and even then only at first and third codon positions, this further supports the observations and conclusions in Chapter 3 that *S. sonnei* shows pattern of nucleotide polymorphisms that are more *E. coli* – like than the rest of the *Shigellae*.

4.4 – Amino Acid Polymorphisms

Given that the patterns of nucleotide change observed at first and third codon positions can be linked to patterns previously observed in chapter 3 at NQ and Q sites respectively, it is interesting to explore the trends in nucleotide change evident at the second codon position, with there being a stronger +AT/+GC bias than at any of the other site classes as well as a trend in the Ti/Tv ratio not observed at the other site classes. The absolutely nondegenerate nature of the second position and the generally radical amino acid changes that result suggest that these unique trends may relate to patterns of selection which are evident at the amino acid level. To that end this section explores the patterns of Amino Acid gain and loss, as well as the metabolic costs associated with those changes in the *Shigellae* and *E. coli*.

4.4.1 – Gain / Loss Biases

Whilst there is substantial variation in the value of the Gain/Loss bias observed (Table 4.4.1a, below), there is a fairly balanced distribution of gainers and losers, with ten amino acids show a net gain greater than +0.1 and eight showing a net loss greater than -0.1. The remaining two amino acids show no negligible bias; Leucine (L) and Valine (V) with mean biases of +0.067 and – 0.006 respectively. The most extreme biases are also relatively evenly split between gainers and losers, with the three strongest gainers showing a mean bias greater than +0.3 and the two strongest losers showing a mean bias greater than -0.3.

Two notably loss-biased amino acids are Alanine (A) and Proline (P) with mean biases of -0.431 and -0.475 respectively, whilst two of the stronger gain-biased amino acids are Cysteine (C) and Isoleucine (I) which have mean biases of +0.604 and +0.389 respectively. Interestingly the more positively biased amino acids tend to be more metabolically costly than the negatively biased amino acids, with those with a mean bias greater than +0.1 having an average cost of 30.7 and those with a mean bias less than -0.1 having an average cost of 23.7, suggesting a relationship between cost and gain/loss bias.

A plot of Gain/Loss bias versus amino acid cost (figure 4.4.1a) shows a significant trend although only when Tryptophan is excluded, this is likely due to the very low number of changes observed which involve tryptophan in all taxa, potentially skewing its G/L bias. The regression statistics with tryptophan are Adj-R^2 = 0 and p = 0.455, with tryptophan excluded they are Adj-R^2 = 0.186 and p = 0.037. So overall there is a strong suggestion that there is a bias towards the gain of more metabolically costly amino acids.

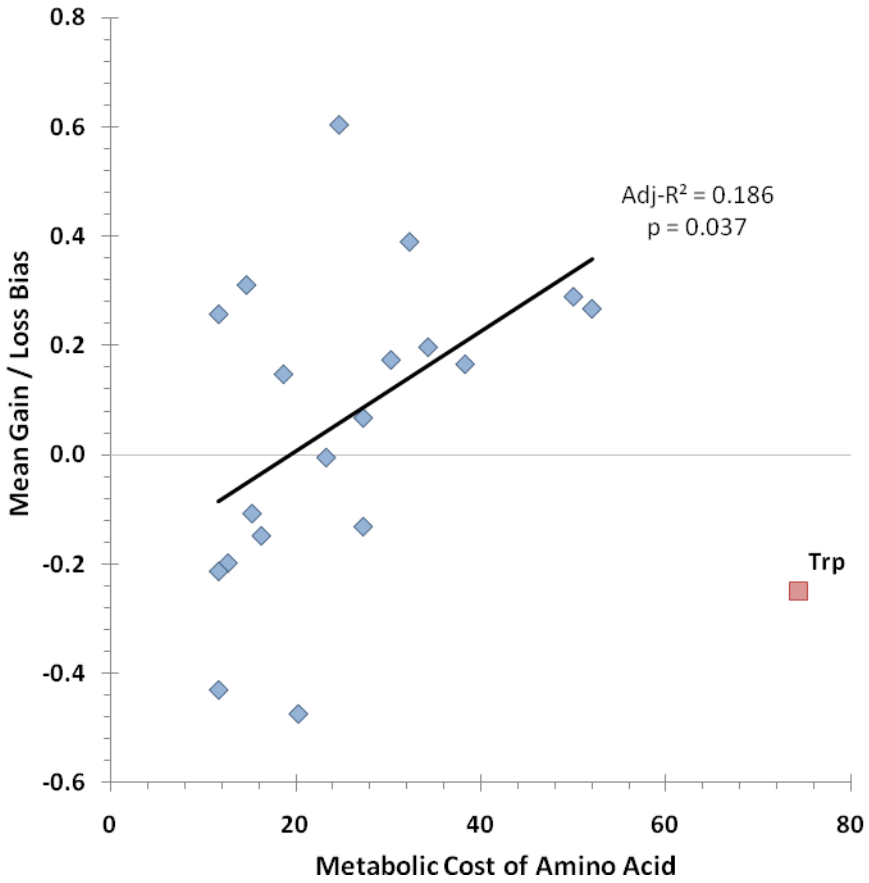


Figure 4.4.1a – A plot of the Mean Gain / Loss Bias observed versus the Metabolic cost of the encoded amino-acid. Tryptophan (W) is shown as a red square and is not included in the regression as very few changes involving W were observed and so the G/L bias is not reliable, additionally this may be due to loss of gene from Trp biosynthetic pathway.

Exploration of the origin of this bias requires examination of the relative abundance of the amino acids within the common proteome of the taxa. A plot of the abundance of the amino acid residues against their metabolic cost reveals a strong significant correlation (Pearson's r = -0.515, p = 0.020) such that more abundant amino acids tend to be less costly (figure 4.4.1b). Additionally a similar plot of the abundance of the amino acid residues against their gain/loss bias reveals a weaker correlation (Pearson's r = -0.407, p = 0.075) such that more abundant amino acids tend to be loss-biased (figure 4.4.1c).

Overall this supports the observation that the more loss biased amino acids are the lower cost amino acids, as a result of being overrepresented in the proteome (relative to their mutational equilibrium frequency) due to selective purging of mildly deleterious changes to more metabolically costly amino acids.

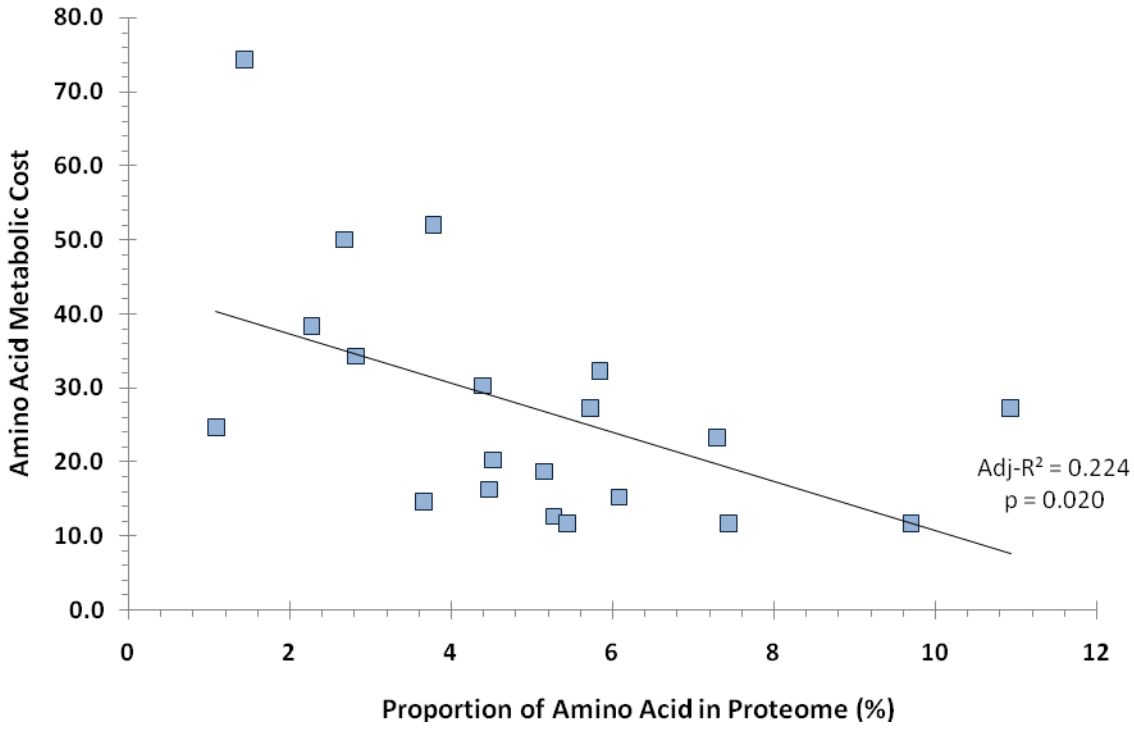


Figure 4.4.1b– Plot of the abundance of an amino acid versus its metabolic cost

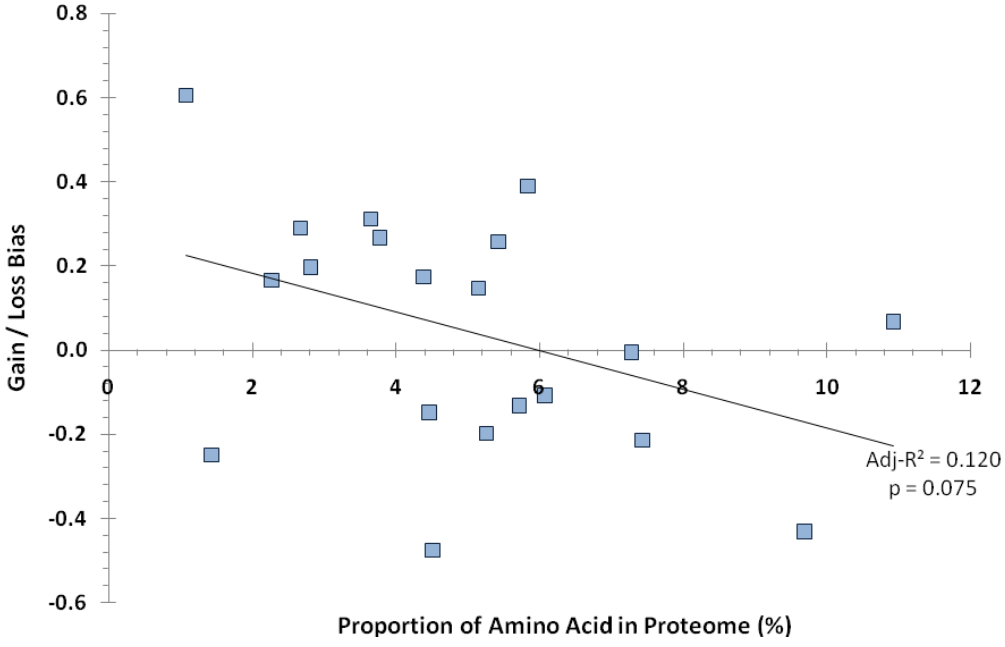


Figure 4.4.1c– Plot of the abundance of an amino acid versus its Gain/Loss Bias

Taxon / Branch ID	Gain / Loss Bias																			
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
EcA	-0.444	0.500	-0.143	-0.216	0.118	-0.323	0.000	0.278	0.315	0.170	0.385	0.534	-0.434	-0.137	-0.122	0.153	-0.033	0.060	0.000	0.212
EcB	-0.473	0.905	-0.303	-0.174	0.077	-0.158	0.273	0.394	0.243	0.040	0.208	0.423	-0.279	-0.180	-0.204	0.244	0.084	0.072	-0.333	0.310
EcC	-0.413	0.360	-0.118	-0.173	0.083	-0.125	0.262	0.379	0.070	0.203	0.278	0.277	-0.600	-0.231	-0.298	0.200	0.145	-0.017	-0.250	0.261
EcD	-0.396	0.487	-0.060	-0.143	0.200	-0.093	0.372	0.260	0.171	0.170	0.188	0.220	-0.426	-0.374	-0.169	0.190	0.105	-0.052	-0.600	0.263
EcE	-0.441	0.923	-0.171	-0.143	0.391	-0.247	0.094	0.422	0.121	0.167	0.192	0.235	-0.608	-0.353	-0.225	0.250	0.070	0.022	0.333	0.586
Sb	-0.457	0.618	-0.306	-0.161	0.185	-0.363	0.327	0.422	0.325	0.319	0.085	0.409	-0.549	-0.138	-0.429	0.364	0.086	0.075	-0.636	0.310
Sd	-0.564	0.811	-0.278	-0.173	0.115	-0.331	0.267	0.340	0.140	0.296	0.149	0.401	-0.481	-0.173	-0.305	0.410	0.151	0.094	-0.586	0.542
Sf	-0.463	0.750	-0.242	-0.244	0.175	-0.333	0.268	0.402	0.083	0.217	0.156	0.399	-0.469	-0.138	-0.312	0.322	0.103	0.094	-0.250	0.576
Ss	-0.592	0.600	-0.157	-0.208	0.176	-0.291	0.151	0.524	0.284	0.057	0.433	0.425	-0.636	-0.171	-0.167	0.409	0.105	0.093	-0.333	0.391
iA	-0.297	0.500	-0.333	0.061	0.143	-0.545	0.263	0.500	0.071	-0.308	0.412	0.257	-0.333	0.000	-0.238	0.250	0.319	-0.137	-0.333	-0.286
iB	-0.581	-0.143	-0.263	0.143	1.000	-0.167	0.200	0.833	0.333	0.111	0.000	0.167	-0.750	-0.400	0.429	0.379	0.214	-0.235	-1.000	-
iC	-0.324	1.000	-0.081	0.000	0.500	-0.067	-0.167	0.286	-0.238	-0.280	0.091	0.333	-0.474	0.200	0.368	0.143	0.219	-0.091	-	0.333
iD	-0.516	1.000	-0.125	-0.167	0.143	-0.143	0.091	0.400	0.400	-0.048	0.091	0.200	-0.333	-0.091	0.077	0.150	0.378	0.022	-	0.333
iE	-0.078	0.143	-0.205	0.081	0.429	0.188	-0.086	0.014	0.106	-0.175	0.083	0.067	-0.280	0.100	-0.256	0.131	0.111	-0.078	1.000	-0.077
Mean	-0.431	0.604	-0.199	-0.108	0.267	-0.214	0.165	0.389	0.173	0.067	0.196	0.310	-0.475	-0.149	-0.132	0.257	0.147	-0.006	-0.249	0.289

Table 4.4.1a – Gain/Loss Bias values (Gains + Losses)/(Gains – Losses) for each Amino-acid, for each taxon/internal branch. With negative values representing a bias towards loss of the amino acid and positive values a bias towards gain. Cells containing a hyphen denote no detection of changes involving that amino acid.

4.4.2 – Metabolic Cost of Amino Acid Changes with Time

Examining the mean cost per amino acid change over time, based upon the extant sequences of the terminal branch taxa and the PAML derived sequences of the internal branches, reveals a significant (Adj-R^2 = 0.371, p = 0.012) trend with time towards a lower cost, as can be seen in figure 4.4.2a (below).

This is consistent with the patterns of nucleotide change observed in Chapter 3 which showed gradual time dependant purging of the more deleterious changes. In the amino acid changes this equates to the purging of the more metabolically costly amino acids in favour of cheaper alternatives. It is worth noting however that the more costly amino acids present in the proteome may be necessary for the structure of a particular protein or play an essential role in its function.

The lack of any clear separation of the *Shigellae* and *E. coli* suggests that the selective constraints upon both groups, in terms of the metabolic cost of proteins, are similar. Given the general traits of reduced purifying selection seen in the *Shigellae* at the nucleotide level (Chapter 3), it is also possible that metabolic cost is more strongly selectively constrained than nucleotide content.

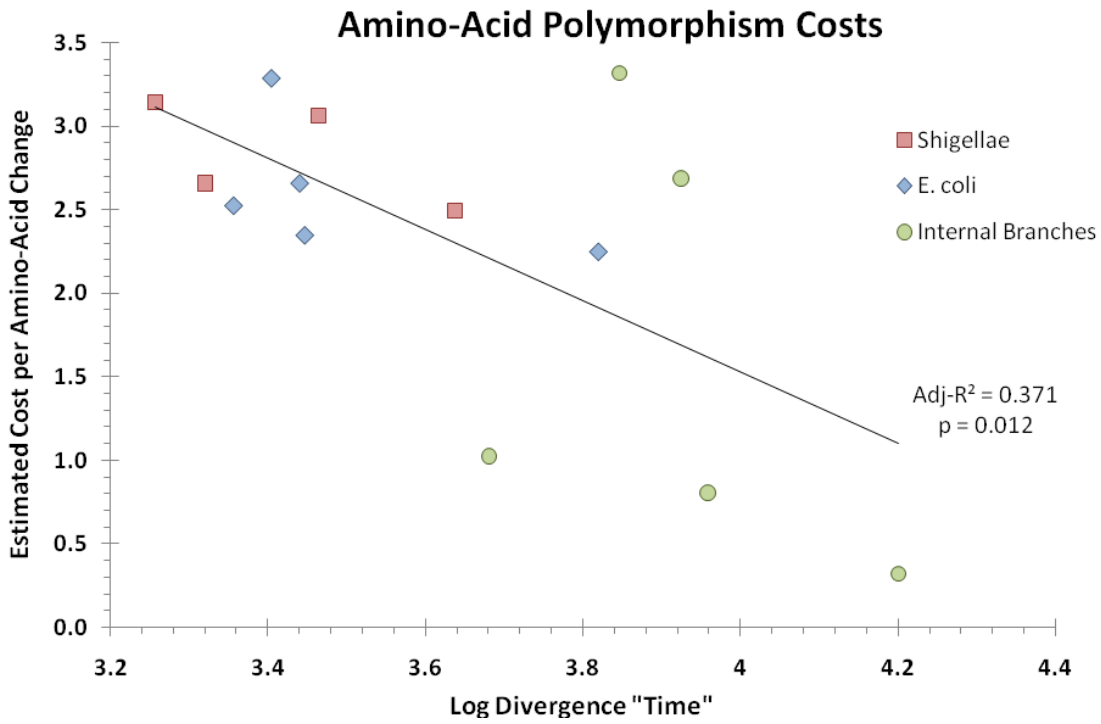
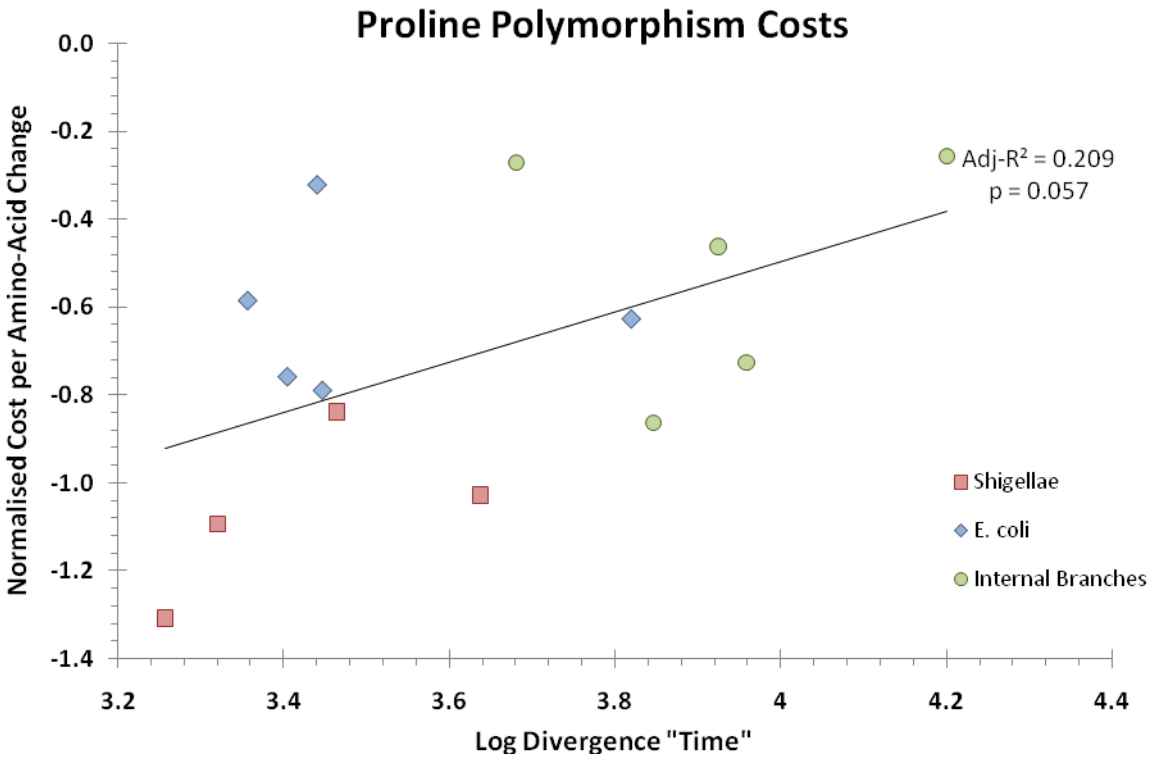


Figure 4.4.2a – A plot of the mean cost per amino acid polymorphism observed in each taxon / internal branch versus time.



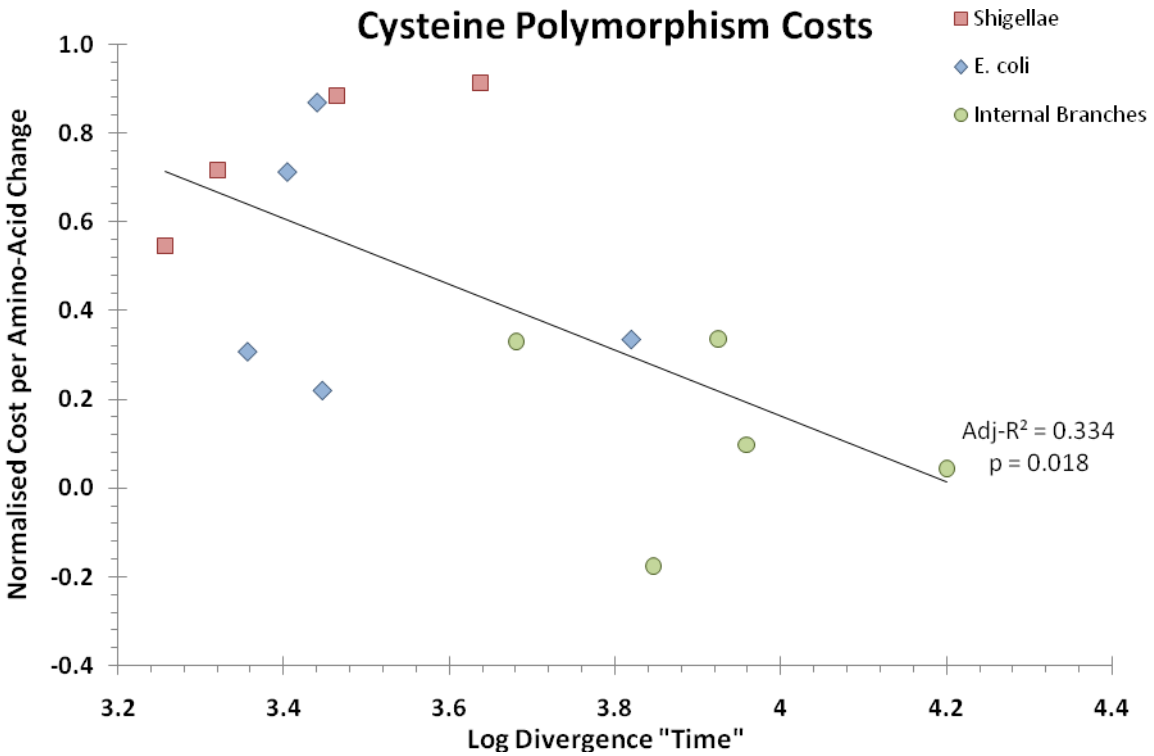


Figure 4.4.2b & c– Plots of the normalised cost per amino acid change for the most loss biased and most gain biased amino acids, Proline & Cysteine respectively.

Cysteine and Proline show the strongest mean Gain/Loss biases, in favour of gain and loss respectively, however there is no significant trend with time for the Gain/Loss bias for either amino acid (Adj-R^2 = 0 for Cys and Adj-R^2 = 0.075 for Pro).

The normalised cost per amino acid change for both Cys and Pro does however show a significant trend with time, with the most gain-biased amino acid (Cysteine) showing a decrease in normalised cost with time (figure 4.4.2c) and the most loss-biased amino acid (Proline) showing an increase in normalised cost with time (figure 4.4.2b). The normalised cost values are such that the sum for all 20 amino acids equals the overall mean cost per amino acid change. This suggests that the decreasing mean overall cost per amino acid change is likely the result of the purging of both losses of more common amino acids and the gains of rarer ones.

Whilst, in figures 4.4.2 a, b &c, there is an apparent separation of the *Shigellae* and *E. coli* only the data for Proline shows a significant difference; p values for t-tests of the residuals to the overall regression line are 0.286, 0.768 and 0.016 for All amino acids, Cysteine and Proline respectively.

4.5 – Summary of Results

- Time dependant purging of 1st and 2nd site SNPs reflects the trends observed in the d*N*/d*S* ratio and the purging of NQ site SNPs in Chapter 3

- +AT/+GC ratio shows consistent patterns across all three codon positions with significant trends towards purging of +AT SNPs with time.

- Ti/Tv ratio shows trends with time reflecting patterns seen in Chapter 3, with the largely synonymous 3rd site showing a trend akin to that at Q sites, and the largely nonsynonymous 1st site showing a trend paralleling that seen at NQ sites. The second codon position showing no convincing trend, likely due to low numbers of observed SNPs.

- Where evident separation of the *Shigellae* and *E. coli* is consistent with reduced purifying selection in the former, as discussed in Chapter 3

- There is a strong and significant correlation between the metabolic cost of an amino acid and its mean Gain/Loss bias across the dataset, with the more costly amino acids showing a Gain bias.

- Additionally both the Gain/Loss bias of an amino acid residue and its metabolic cost show a correlation, the latter more strongly than the former, such that more abundant amino acids are both cheaper and more loss biased.

- The mean cost per amino acid change shows a downward trend with time reflecting the purging of more metabolically costly amino acid changes

4.6 – Discussion

4.6.1 – Relative abundance of SNPs at each codon position

Overall the proportion of SNPs at any given codon position is in line with expected values given the degeneracy of the third codon position evident from the genetic code, in addition the trend with time also reflects the known time dependency of the purging of more deleterious SNPs (Rocha, Maynard Smith et al. 2006) in this instance the purging of SNPs at the first and second codon positions. Those SNPs at the first position being relatively more conservative than those at the second position, which is reflected in their presence at a slightly higher level in the sequence data.

4.6.2 – Metric ratio variation with codon position

The +AT/+GC ratio shows a consistent bias towards AT enriching mutations, in line with known biases in the mutation processes in *E. coli* (Halliday and Glickman 1991; Schaaper and Dunn 1991), namely the predominance of the deamination of Methyl-Cytosine and the subsequent correction of the abnormal uracil-like base to Thymine. Both the first and third codon positions show similar levels of AT enrichment, however as mentioned there are differences in line with their base composition, in that the more AT poor first site experiences a slightly but significantly higher +AT/+GC. This will be the result of a mutation bias towards restoring the mutational equilibrium base content which, given the aforementioned inherent bias towards T enrichment, would likely be an AT rich genome.

The second codon position presents somewhat of a conundrum with respect to the values of the +AT/+GC ratio; it is far higher than both the first and third codon positions (mean ratios of 2.66 versus 2.11 and 1.83 respectively), which may well reflect and increased mutation bias towards AT however this is in a background of AT rich sequence (approximately 59%). This suggests that either the mutation bias towards AT is extremely high at second positions, but not first and third, or that there is a strong selective pressure towards AT enrichment. The former is unlikely given the high AT content of second positions leaving the possibility that there is strong selective pressure to maintain a high AT content at the second codon position. Given that a high AT content at the second

codon position has been previously observed across several bacterial genera (Majumdar, Gupta et al. 1999), it is likely that it is the selective maintenance of hydrophobic amino acid residues at abundances above their mutational equilibrium which has resulted in a bias towards AT richer sequences at second codon positions. Additionally the paucity of SNPs identified at the second codon position means that the trends observed at these sites are less strongly supported than those at first and third positions.

The Ti/Tv ratio shows a bias across all sites towards Transitions, which are largely less selectively costly and mechanistically more common than transversions. Whilst the first and third positions again show similar trends and show no statistically significant difference, the patterns observed are most easily explained by two different processes, although both are products of the composition of the genetic code. The first position bias results from the relatively conservative nature of transitions as compared to transversions, whilst the third position mean Ti/Tv ratio of 2.933 can be explained by a selective bias resulting from the presence of twofold degenerate sites at the third position which are synonymously linked exclusively by transitions (Ti).

The second codon position has values of the Ti/Tv ratio which reflects the previously described mutational bias towards transitions (Collins and Jukes 1994). The values observed range from 1.5 to 2.5, with a mean of 1.7, this is again in line with the absolute nonsynonymous nature of changes at these sites resulting in little selective difference between transitions and transversions, and agrees with the findings of Collins and Jukes, that where amino acid changes are caused by a single nucleotide change there is a strong bias towards transitions.

4.6.3 – Metric Ratios at Codon Positions over Time

Examination of the +AT/+GC ratio over time reveals that the general trend observed in Chapter 3, holds – there is a gradual time dependant purging of AT enriching polymorphisms at all three codon positions. In each case the trend is both strong and significant, with both first and third codon positions showing a clear trend towards a minimal observed value of approximately one (0.981 and 1.013 for first and third positions

respectively) observed at the iE internal branch – the phylogenetically deepest, and consequently 'oldest' data point. The trend towards parity of the +AT/+GC ratio hints at the possibility that the selective bias is towards maintenance of the AT content of the genome, however as tempting as it is to speculate on the selective equilibrium the dataset provides insufficient evidence to test or support a conclusion.

The second codon position shows the same time dependant purging of AT enriching SNPs as the first and third codon positions, whilst it shows a trend roughly parallel to that of the first position, there is clearly a faster rate of decrease in +AT/+GC at the first and second positions than at the third position. This is expected, however it would also be expected that the higher selective constraint on the second position would result in the rate of purging of deleterious SNPs being highest at those sites.

There are potentially complicating factors other than simply selection based upon degeneracy as observed in the genetic code, for example; it has been noted that there is a pattern of predominance of either A or T at second positions which correlates with secondary structure (Gupta, Majumdar et al. 2000) the former being more prevalent in Alpha helices and the latter being more prevalent in Beta sheets. Consequently the high +AT/+GC ratio may reflect the favouring of the desired bias (towards specifically A or T) in order to maintain the secondary structure of the encoded protein, whilst the downward trend is indicative of the selective purging of the undesired A or Ts where appropriate, the signal being highly obfuscated by the examination of 2098 concatenated orthologues rather than specific examination of sequences encoding known secondary structures.

The lack of consistency observed in the time dependant trends of the Ti/Tv ratio at each codon position is likely related to their degeneracy coupled with the selective consequences of changes at each site. Only the first codon position shows any significant trend with time – the gradual purging of Transversions, which can be explained in the light of the relatively non conservative nature of transversions (Zhang 2000).

Whilst there are a slightly higher number of fourfold degenerate sites than twofold degenerate sites at the third codon position (32 vs 24 respectively), the corresponding bias towards selective equality of transitions and transversions (where they occur at fourfold degenerate sites) is insufficient to explain the absence of a time dependant trend in the Ti/Tv ratio at this position. It would be expected that given the presence of two-fold degenerate sites, that there would be a selective difference between transitions and transversions and consequently a time dependence of the selective purging of the transversions. There is little evidence from the data as to why there is no trend at this position and the results observed provide no insight into any potential explanations.

The second codon position also shows no significant change in the Ti/Tv ratio with time due again to lack of selective differentiation between transitions and transversions, however not due to degeneracy but due to the absolute lack of degeneracy and the largely non conservative nature of amino-acid changes resulting from second codon position changes. Additionally the paucity of SNPs observed at the second position means that there may be trends present but there simply isn't enough resolution in the dataset to observe them.

4.6.4 – Patterns in Gain/Loss Bias of Amino Acids

As can be seen in figure 4.4.1a, there is a clear relationship between the metabolic cost of an amino acid and the Gain/Loss ratio observed, such that the more metabolically costly an amino acid is the more likely it is to be gained, based upon the changes observed in the data set. This at first is indicative of the "universal trend of amino acid gain and loss" originally posited by Jordan et al (Jordan, Kondrashov et al. 2005), however as with that case it must be borne in mind that the comparisons here are between closely phylogenetically related species or strains and so are more likely to show patterns akin to the mutation bias than the selective one (Hurst, Feil et al. 2006; McDonald 2006). Under a mutation biased pattern of gains and losses those amino acids which show a strong gain bias would be expected to be underrepresented in the proteome and those showing a strong loss bias, overrepresented (figure 4.4.1b). The lack of significance likely reflects the interplay of both selection and mutation on the gain/loss bias of any given amino acid

as each will have differing levels of selective benefit/cost, resulting in some amino acids whose abundance is dictated predominantly by mutational biases and others whose abundance is dictated by more selective biases.

However, there is a clear and significant trend with the mean cost of amino acid polymorphisms towards the gradual purging of more costly changes (apparent from the decreasing average cost). The actual values of the mean cost per change are always greater than zero, suggesting that, on average change is costly. In this situation it would be expected that the metabolically cheaper amino acids are over represented in the proteome, and the more expensive underrepresented, a correlation of percent abundance with cost reveals that the two variables are significantly correlated (Pearson's r = -0.515, Adj-R^2 = 0.224, p = 0.02).

Overall this suggests that the patterns of amino acid gain and loss, at least from mutational perspective, are a reaction to the selective pressures acting upon the genome to minimise the metabolic cost of the proteome. The high cost amino acids are purged, where their loss does not itself incur a fitness cost, creating a compositional imbalance in the proteome, resulting in a mutational bias to revert that imbalance. Indeed it has been shown that selective pressures correlating to metabolic synthesis cost have affected amino acid usage in all three domains of life (Swire 2007).

Examination of the trends associated with the normalised mean cost per AA change of the most positively and negatively G/L biased amino acids further reinforces this. The positively biased (gainer) Cysteine shows a time dependant purging of changes from 'costly' (postitive cost values) towards a contribution to the overall mean of zero (restoring it's underrepresentation), whilst the negatively biased Proline shows a gradual time dependant purging of changes from energetic gains (negative cost values) towards zero (restoring its overrepresentation).

4.6.5 – Separation of the *Shigellae* and *E. coli*

The nucleotide patterns at codon positions show some separation of the *E. coli* and *Shigellae*, the percentage of SNPs at third site shows the greatest separation when considering codon position distribution of the SNPs, the lower proportion of SNPs at the more nonsynonymous position being congruent with the reduced purifying selection inherent to the lifestyle niche occupied by the *Shigellae*. However the ratio of the number of SNPs at 2nd positions to the number of SNPs at 1st positions also provides an interesting separation of the *E.coli* and *Shigellae* (figure 4.2.2b). The *E. coli* show a strong and significant trend representing the preferential purging of 2nd position SNPs over 1st position SNPs, whereas the *Shigellae* show no trend, reflecting reduced purifying selection and thus an inability to purge 2nd position SNPs in preference to 1st position SNPs.

The first codon position (and less so third) shows a much higher proportion of AT enriching SNPs in the *Shigellae* again supporting the observed trend for the *Shigellae* to have a higher level of mildly deleterious changes in their genome than the *E. coli* due to reduced purifying selection, additionally it has been shown (Singer and Hickey 2000) that more AT rich genomes have a higher proportion of more costly amino acids (F I N K Y, mean cost = 35.6) and GC rich genomes have a higher proportion of cheaper amino acids (G A R P, mean cost – 17.8) further underlining the deleterious nature of AT enriching SNPs. There is also an observably greater proportion of the more deleterious transversions at first codon positions, also in line with reduced purifying selection in the *Shigellae*.

Examination of differences in the amino acid trends between the *E. coli* and *Shigellae* reveal no significant difference in the G/L biases (paired t-test of mean G/L biases for each amino acid; $p = 0.968$). Whilst there appears to be differences in the time-dependant trends of cost of amino acid change/replacement between the *E. coli* and *Shigellae* only the trend in Normalised cost per Proline change shows significant difference (as determined by a t-test of the residuals to the regression line; $p = 0.013$) the

Shigellae showing a more negative cost, representing lesser purging of Proline losses, again consistent with reduced purifying selection.

Overall the separation of *E. coli* and *Shigellae* by the pattern of nucleotide polymorphisms observed at the codon positions is consistent with earlier conclusions in Chapter 3 as regards the likely evolutionary mechanisms involved – reduced purifying selection in the *Shigellae* as a consequence of their relatively recent niche specialisation and consequent reduction in effective population size. The lack of consistently significant separation from *E. coli* by differences in the amino acid metrics however suggests that either the amino acid sequence of the 'core' genome of the *E. coli* / *Shigellae* shows a higher level of selective constraint than the nucleotide sequence or that the effects of reduced selection potentially require longer periods of time to become evident at the amino acid level.

In general however wherever there is a time dependence of a metric there is a degree of apparent separation of the *Shigellae* and *E. coli*, strongly supporting the earlier conclusions both in this Chapter and in Chapter 3 that the differences observed are due to reduced purifying selection in *Shigellae* as a consequence of their recent lifestyle change and concomitant reduction of effective population size.

Chapter 5 – Patterns of Nucleotide Substitution around the Genome

5.1 - Introduction

5.1.1 – Bacterial Genome Organisation

In the vast majority of cases the bacterial chromosome is circular and shows a highly conserved set of features, independent of the genetic content. The system of replication of the bacterial chromosome divides it into two replichores, each separated by the origin of replication (Ori) at one end and the replication termination site (*dif*) at the other; these are essentially equal in size and bring with them biases and patterns of organisation to the chromosome (figure 5.1.1a).

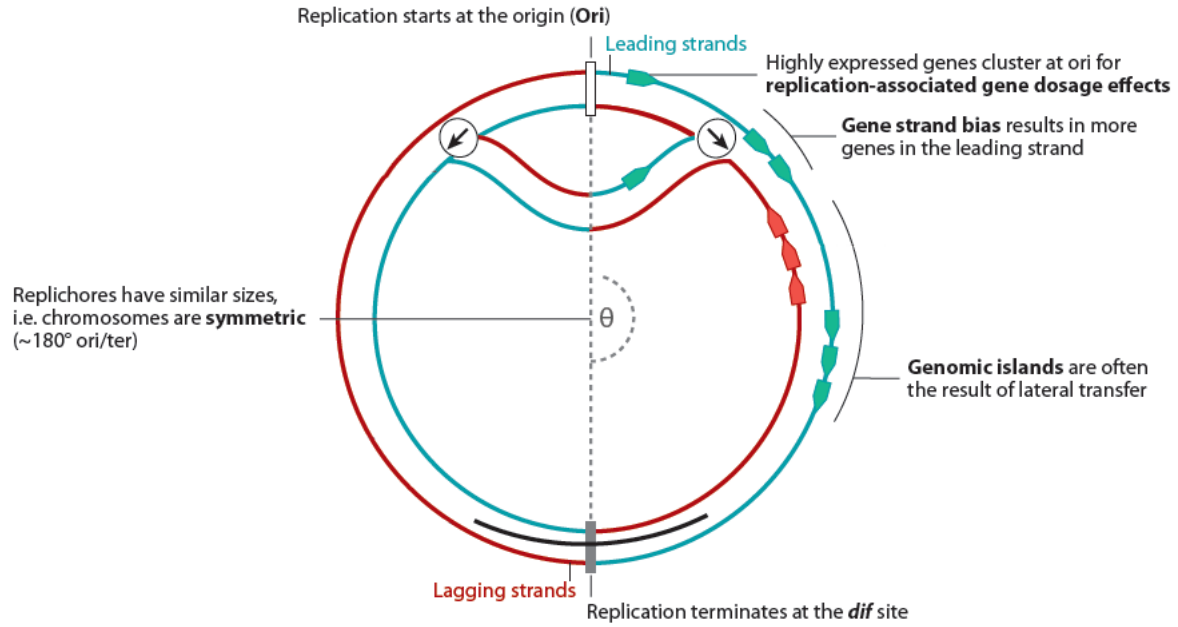


Figure 5.1.1a – Diagrammatic structure of a typical bacterial chromosome, showing some of the key features of the organisation of bacterial chromosomes. Adapted from (Rocha 2008).

The existence of a single origin of replication and the resulting replication of half the genome in a single run by a replication fork means that it can take a significant amount of time (in terms of the bacterial life cycle) to replicate the entire genome. Consequently fast replicating organisms will possess multiple nested replication forks, that is the genome with initiate a new round of replication before the previous one has finished, as is the case in *E. coli*.

The number of nested initiations of replication can be estimated by calculating the ratio of the time taken to replicate the genome to the time between cell divisions. This ratio (R) can be close to zero (in the case of bacteria with very long generation times), approximate to one (in bacteria where the generation time and genome replication time are approximately equal) or greatly exceed one, as is the case in rapidly dividing bacteria. This nested set of replication forks ensures that the daughter cells of a given bacterium receive an already partly replicated genome. Another aspect of this is that genes close to the origin are in any given cell approximately 2^R times more abundant than those at or near the terminus (Cooper and Helmstetter 1968). In the case of *E. coli* with a time between cell divisions of approximately 20 minutes (under optimal conditions) the time taken to completely replicate the chromosome is a minimum of 33-34 minutes (Bipatnath, Dennis et al. 1998), and so copies of genes near the origin are on the order of four to eight times more abundant than those near the terminus, as in the example below (figure 5.1.1b).

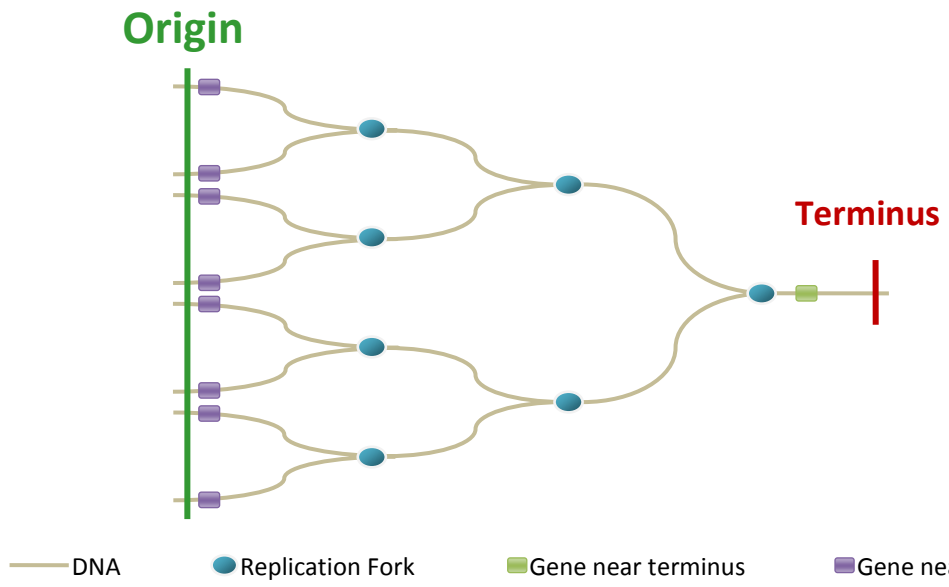


Figure 5.1.1b – Representation of the staggering of multiple rounds of replication within a single replichore, each set of vertically aligned replication forks corresponding to one round, demonstrating the greater copy-number of genes near the origin of replication. Representing an R value of 3.

In the absence of moderation by gene regulation, this results in greater levels of expression for genes near the origin (Sousa, de Lorenzo et al. 1997) which is of potential evolutionary benefit to the organism. This is especially the case in rapidly replicating bacteria where genes associated with replication have been observed to be clustered

near the origin (Couturier and Rocha 2006), it has also been shown that inversions which move genes near the origin further away have the potential to affect optimal growth rates and viability of the cell (Campo, Dias et al. 2004).

The nature of replication forks creates biases between the leading and lagging DNA strands in the location of classes of genes. There is a preponderance of essential genes on the leading strand (Rocha and Danchin 2003), which is likely a consequence of the need to avoid head-on collisions between the replication fork and the transcription complex. Such a collision has the potential to halt the replication fork which not only renders the locus unavailable for a considerable amount of time but also can result in higher mutational load on such loci due to rescue of the replication fork by homologous recombination (Mirkin and Mirkin 2005). Consequently, genes on the leading strand are less likely to halt the replication machinery and are more likely to complete transcription before being dislodged, resulting in fewer potentially harmful partial transcripts or peptides.

There are also variations in mutational patterns observable around the genome; genes closer to the terminus of replication (*dif* site) show levels of divergence approximately twofold higher, than genes near the origin, between *E. coli* and *Salmonella typhimurium* specifically when measured using d*S* as a metric of divergence (Sharp, Shields et al. 1989). Mira and Ochman (2002) confirm the observation made by Sharp et al but at a slightly lower level of 1.5 fold, they also detect a trend of increasing transitions and transversions with distance from the origin the latter increasing more rapidly with distance (where an effect was observed) as well as an effect of greater difference in GC composition of homologous genes with distance from the origin. Given that highly expressed genes involved with essential cell processes (transcription and translation) are located nearer the origin it is possible that the trend is a result of a gradient of selective constraint on the genes, with the less essential and therefore less constrained genes being further from the origin.

5.1.2 – Horizontal Gene Transfer and its Detection

Horizontal or Lateral gene transfer (HGT or LGT) is one of the primary ways in which bacteria can acquire new additions to their genetic toolsets allowing a species or bacterium access to a new ecological niche, enable it to better exploit its current niche or adapt to a change in its environment. There are three main modes by which the new genetic information is acquired; Uptake from the environment (Transformation), Introduction by a virus (Transduction) or Bacterial genetic transfer (Conjugation). In all three scenarios the incoming genetic material is integrated into the host genome via homologous recombination, in a process akin to that of double stranded DNA break repair, which requires that there be a region of high similarity between the new DNA and the host chromosome.

There are many methods of detecting HGT all of them rely on the principal that the incoming DNA sequence will bring with it sequence traits and features more readily associated to its genome of origin than to the new host genome. One of the more readily observable of these is nucleotide bias, such as GC content. If the transferred gene(s) have originated from a distantly related species then they are likely to have a different bias in nucleotide composition to the host genome. In addition to this, features such as codon bias or preference may also be different in the newly acquired genes, however use of measures such as these is prone to false positives as selective constraint on essential genes may, through patterns of amino-acid use, endow them with 'abnormal' nucleotide signatures.

Additionally the variation in phylogenetic relationship between aligned genomic sequences along its length can be used to infer regions of the genome which show 'abnormal' relationships. More recently methods based on examination of changes in the patterns and rates of nucleotide change have been used (Hamady, Betterton et al. 2006), which have been shown to have greater accuracy than methods relying solely on the base composition or inherent sequence biases whilst at the same time avoiding the computational intensity of estimation of multiple phylogenetic relationships.

5.1.3 – Aims & Conclusions

Previous examination of nucleotide substitution patterns has shown significant variation between closely related genomes. Here I aim to examine how those nucleotide substitution patterns vary, via the use of nucleotide site-type distribution and the examination of previously derived metric ratios, according to genome position.

I observe several biases in several metrics and/or genomic features which show clear trends with respect to distance from the origin (AT content, Proportion of SNPs at 3rd or NQ sites & Ti/Tv ratio) as well as some that do not (SNP density & +AT/+GC ratio). Additionally I observe a region in EcC which displays strong biases in many of the ratios consistent with increased evolution.

5.2 – Dataset Synteny

The figure below shows the chromosomal rearrangements of the *Shigellae* genomes in relation to the nucleotide sequence of *E. coli* MG1655 (EcA). As can be seen below there is a great deal of rearrangement in all four *Shigellae*, primarily centred on the terminus and, to a lesser extent, the origin. There is however clear variation in the size and location of rearrangements evident between each *Shigella* and EcA, with *S. dysenteriae* and *S. boydii* showing a large number of smaller rearrangements (cyan in figure 5.2a) and *S. flexneri* showing a bias towards larger rearrangements. The levels of rearrangement are likely the result of a combination of factors, notably the divergence time of each sequence from EcA and artefacts due to the disparate *E. coli* backgrounds from which each *Shigella* will have arisen.

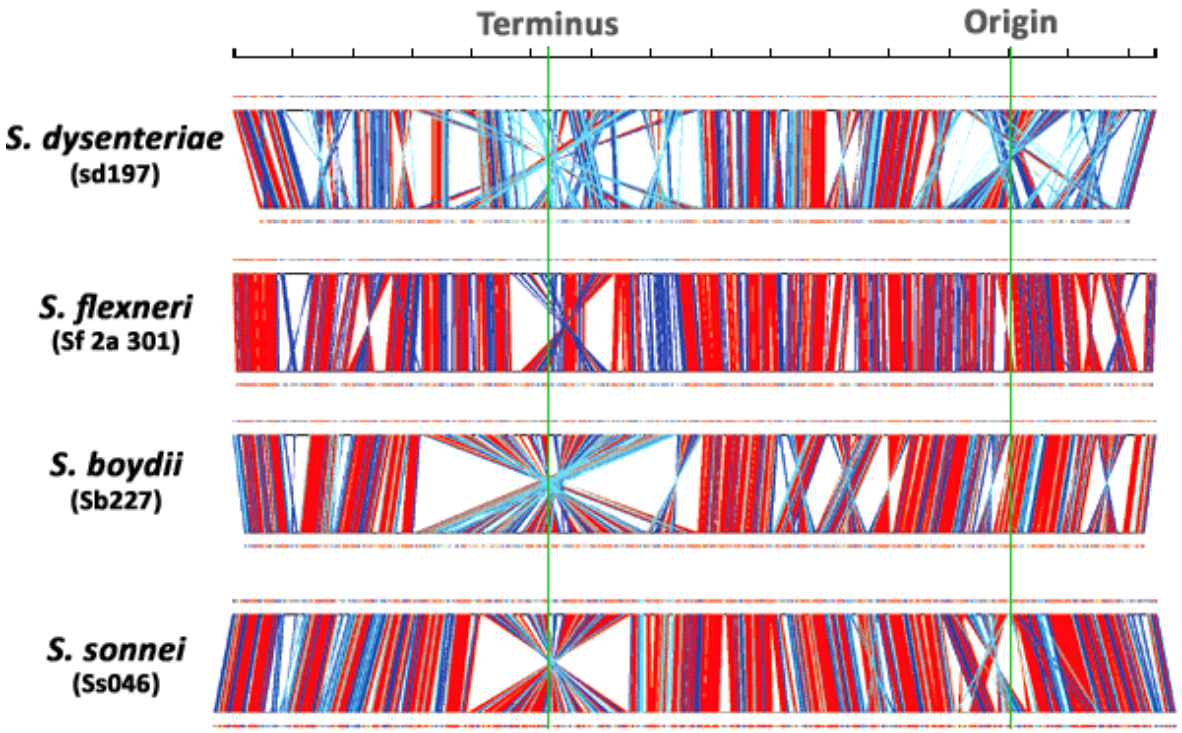


Figure 5.2a – A schematic representation of the chromosomal rearrangements present in each of the *Shigellae* strains used as compared to *E. coli* K-12 MG1655, shown at the top of each comparison. Colour code donates maximal length of the paired segments: red, >10kb; blue, 5~10kb; cyan, 1~5kb. Adapted from Yang et al (2005).

It is worth noting however that the collection of orthologous genes will exclude many of these rearrangements either through the exclusion of non homologous sequences or through the methodology used to conservatively identifying orthologues – any putative

orthologues which fall outside of blocks of conserved synteny are excluded, as they are likely to be false positives.

Examination of the conservation of the gene order between the genomic sequence and the order of the orthologues shows that all the *E. coli* sequences are highly syntenic. The only rearrangement being in *E. coli* UTI89, which is of the two more distantly related of the five *E. coli*, this break with synteny is small however and represents no overall change in the distance of these genes from the origin (figure 5.2b).

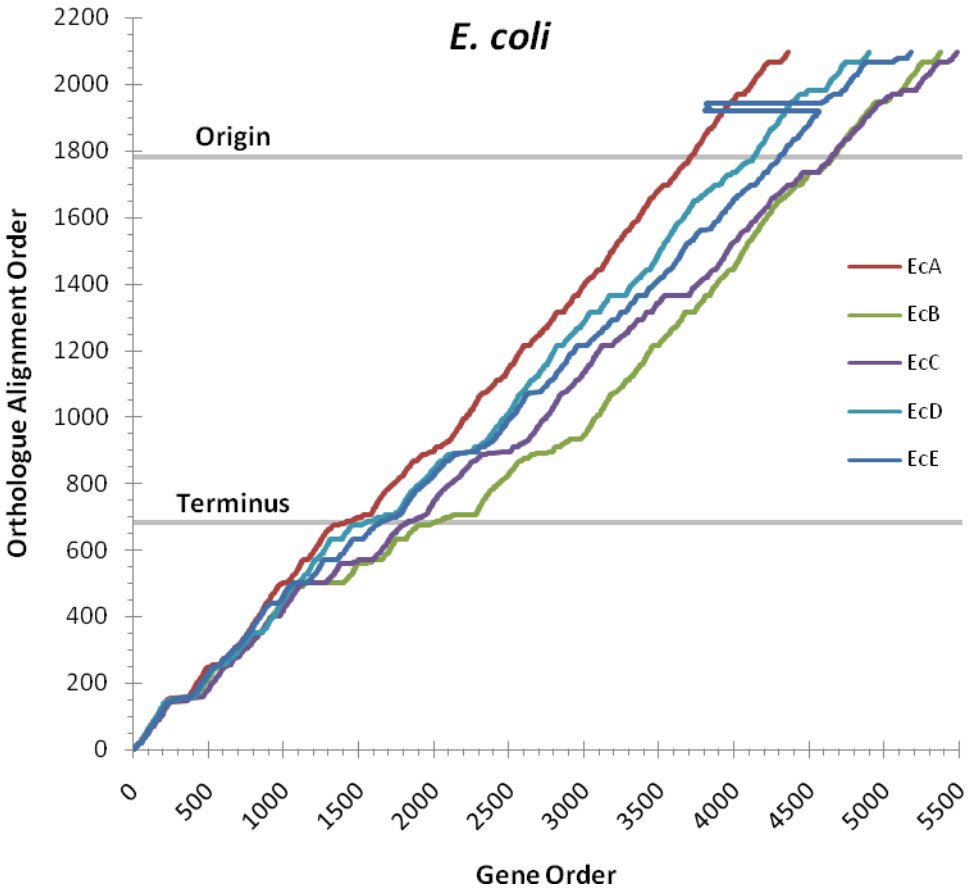


Figure 5.2b – A plot of position in the genome against the position in the orthologous alignments for all the *E. coli* used. Origin location determined via identification of flanking orthologues. Terminus location approximated using information from Niki and Hiraga (1998)

A similar examination of the syntenic conservation in the *Shigellae* reveals large scale inversions and synteny breaks, as with the genome sequence comparison these are centred primarily on the terminus and are inversions. In the case of *S. dysenteriae* the pattern is consistent with multiple inversions centred on the terminus creating a mosaic effect of syntenic and non syntenic sequences; overall there is minimal change in the distance of the genes involved in any of the *Shigellae* from the terminus (figures 5.2c&d).

Aside from the terminus centred inversions there are several smaller inversions between the origin and terminus, most notably in *S. boydii* however these are relatively small and being roughly equidistant from the origin and terminus again have minimal effect on distance to either. Finally there is a moderately sized inversion roughly centred on the origin in *S. dysenteriae*, as with previous inversions this has minimal effect on the distance of any given gene from the origin.

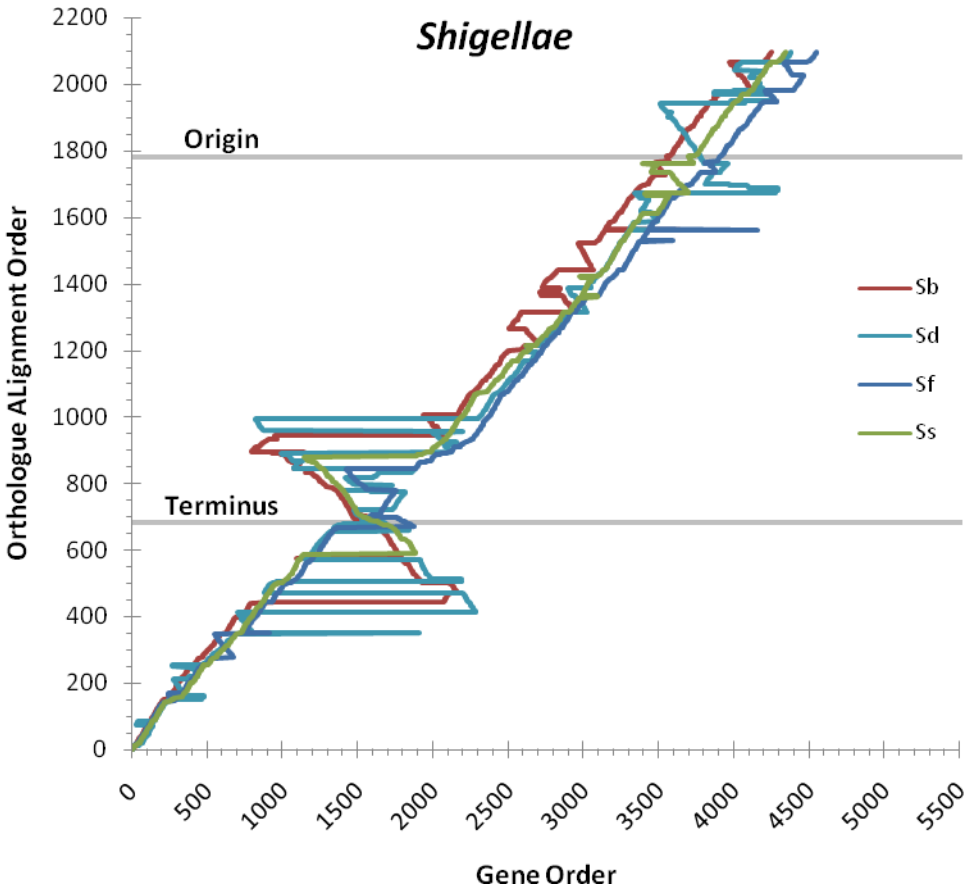


Figure 5.2c – A plot of position in the genome against the position in the orthologous alignments for all the *Shigellae* used. Origin location determined via identification of flanking orthologues. Terminus location approximated using information from Niki and Hiraga (1998)

Overall, whilst the synteny is not absolutely conserved between the *E. coli* and the *Shigellae* there are no changes which significantly alter the distribution of genes relative to the origin or terminus, as is evident from the comparison of distances from the origin in the *Shigellae* (figure 5.2d), and so trends in nucleotide substitution biases related to such distances should be largely unaffected.

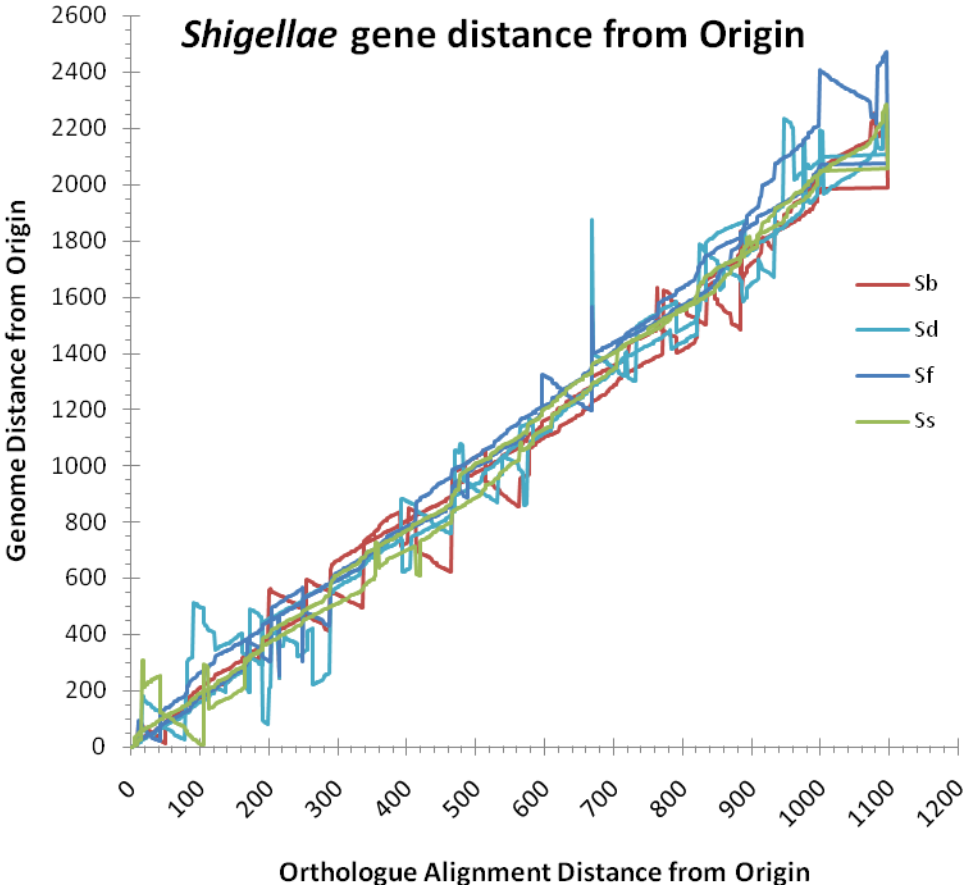


Figure 5.2d – A plot of distance from the origin in both the orthologue alignment and the genomes for all the *Shigellae* used. Origin location determined via identification of flanking orthologues.

5.3 – Nucleotide Bias along the Aligned Core Genome

Sliding window analysis of the AT content of the genomes shows that is a clear trend in the variation of the GC content with genome position, with a strong bias towards more AT rich sequences around the terminus and a lesser but clear bias towards more GC rich sequences around the origin, as is evident in figures 5.3a and b. There is little differentiation between the mean of the *E. coli* windows and the mean of the *Shigellae* windows, where differences are observed there is a tendency for the *E. coli* values to be more 'moderate' that is the *E. coli* AT content is typically closer to the genome AT content than the *Shigellae*. However, this is not an artefact of the overall AT content of the genomes – they all fall between 46.9 and 47.1%.

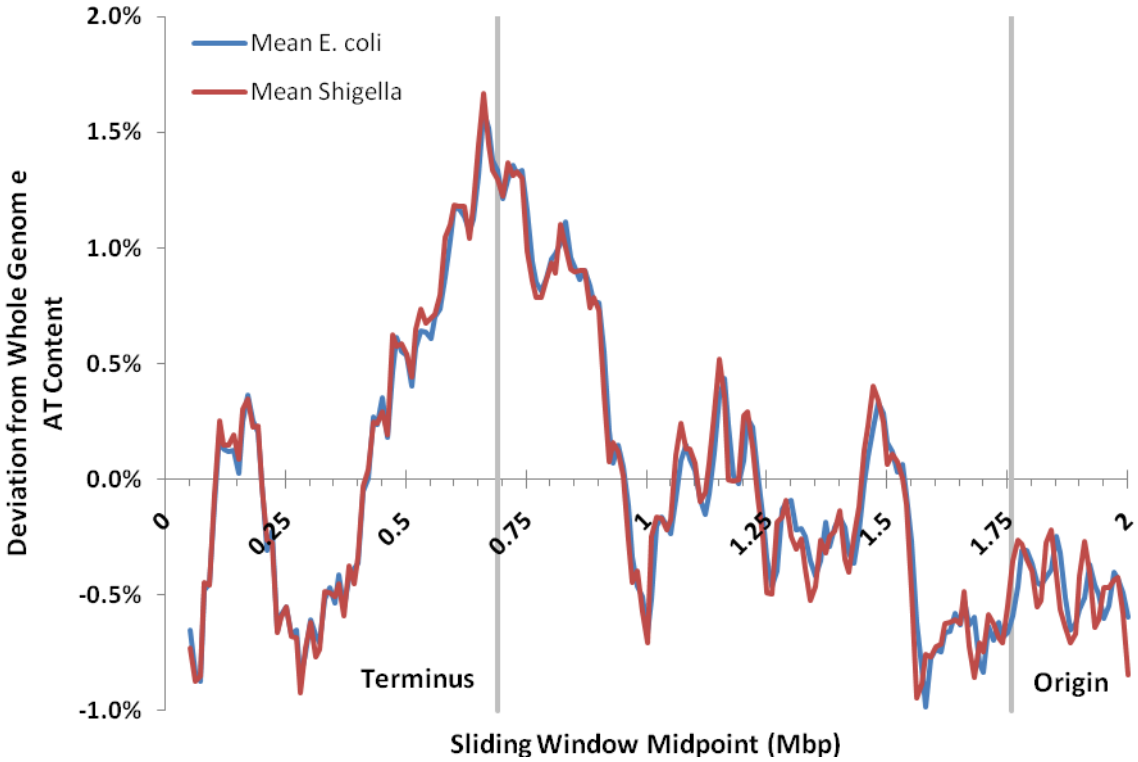


Figure 5.3a – Deviation from the genomic mean AT content against midpoint of the sliding window. Using a window of 100Kbp with a step of 10Kbp.

An examination of the GC and AT skews along the alignment revealed no clear trends with regards to genome location, likely a result of the dataset containing both leading and lagging strand genes, which would mask the strand dependant patterns of base skew, alternatively there may be no significant strand bias in the *E. coli* genomes used.

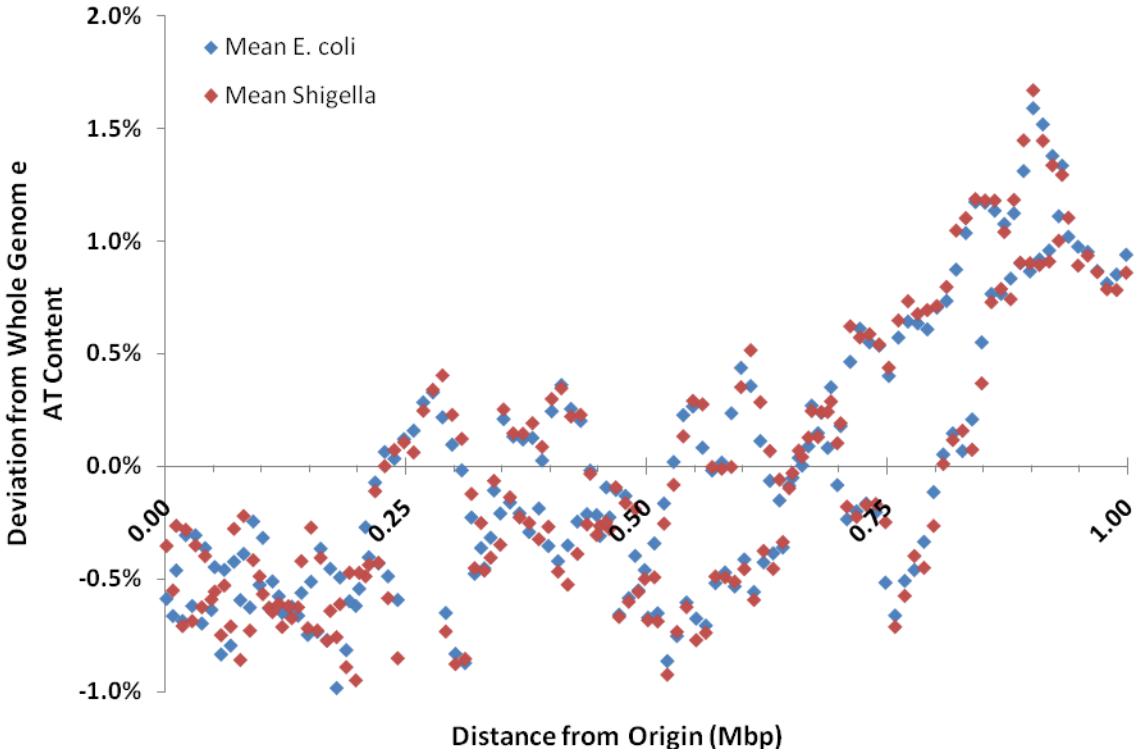


Figure 5.3b – Deviation from the genomic mean AT content as calculated in figure 5.3a, plotted against the distance from the Origin.

5.4 – SNP Distribution

5.4.1 – Density of SNPs

SNP density fluctuates along the alignment for all the sequences analysed; however there are several key areas identified which show strong deviations from the genome mean SNP densities. A strong deviation was defined as plus or minus 1.645 times the standard deviation of the differences observed in a given taxon, and ensures that the 'strong' regions are the top (and bottom) 5% of the differences observed, i.e. the extreme 10% of values, within each taxon.

EcD shows two main regions that are not shared with any of the other genomes; both with lower SNP density at 1.3 – 1.4Mbp and 1.95 – 0.1Mbp. EcB shows the smallest 'strong' deviations from mean SNP density, with most regions being only one window wide (100Kbp). Most other taxa show several multi-window regions ranging from a 1 SNP per Kbp deviation (e.g. Sb from 0.15 - 0.25Mbp) to 4 SNPs per Kbp deviations (EcA 1.65 – 1.75 Mbp & EcD 1.95 – 0.1 Mbp) as can be seen below in figure 5.4.1a.

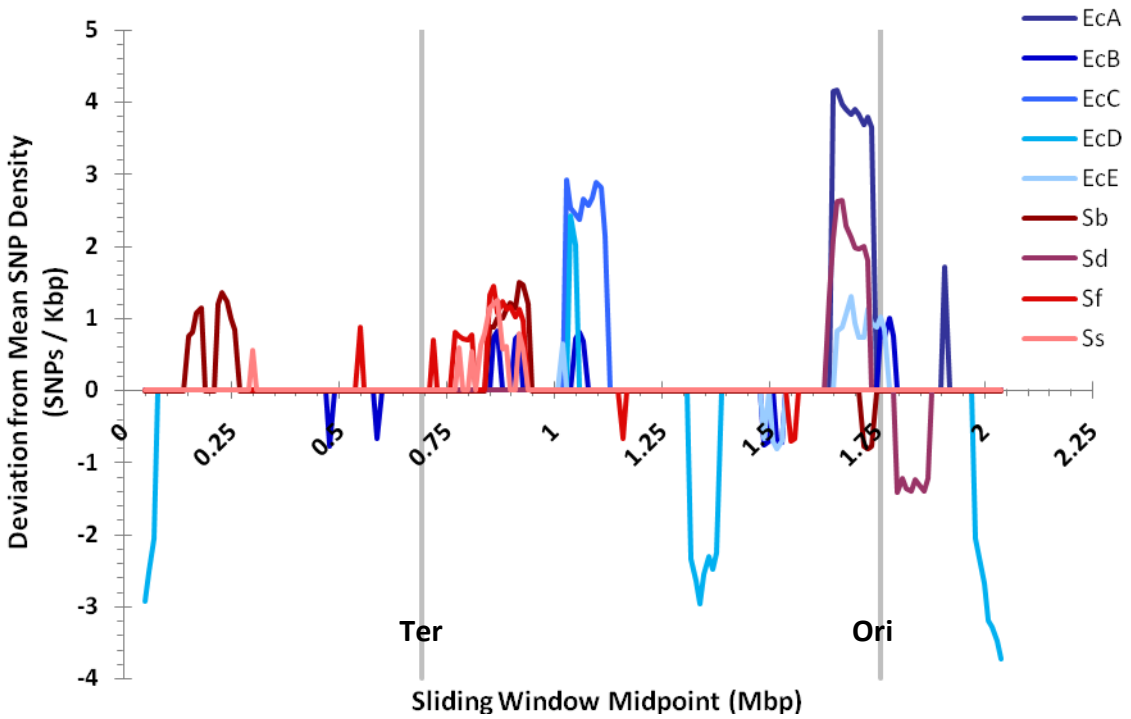


Figure 5.4.1a – Deviation from the genomic mean SNP density, only values of 'strong' differences corresponding to 1.645 standard deviations above or below the mean of all differences are shown.

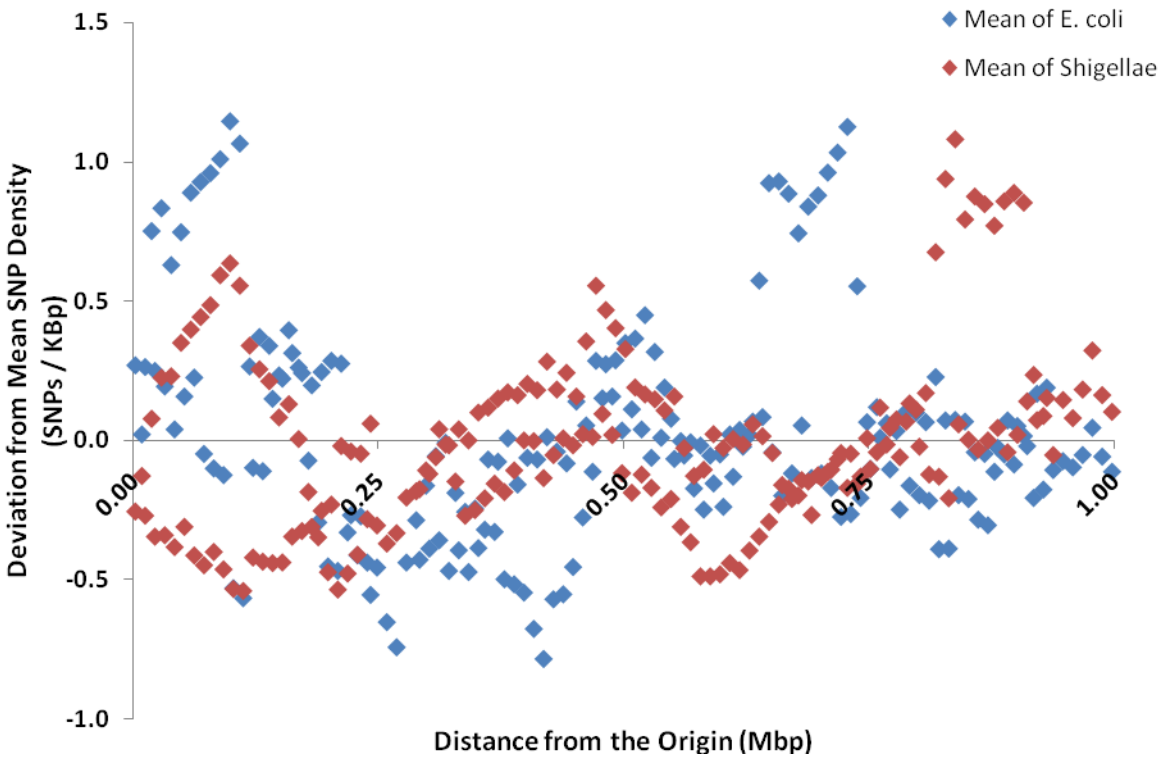


Figure 5.4.1b – Deviation from the genomic mean SNP density against distance from the origin.

Whilst SNP density variation hints at variations in patterns of polymorphism around the genome, it shows no clear pattern or trend with respect to the distance from the origin and also reveals little in the way of detail (figure 5.4.1b). However based upon patterns observed in both Chapter 3 and Chapter 4, the distributions of SNPs between nucletoide site types and the values of the normalised ratios of SNP types could potentially provide further insight into the the patterns of selection and evolution around the genome.

5.4.2 – Codon Position Variation

From figure 5.4.2a, and table 5.4.2a, it is clear that the reduced bias observed in the *Shigellae* in earlier chapters towards a lower proportion of 3rd site SNPs is common along the entire length of the alignment. Both the *Shigellae* and the *E. coli* show similar trends with genome position, generally with a greater proportion of SNPs at the third codon position towards the origin than around the terminus.

This difference is even more apparent when visualising the deviation from the mean proportion of SNPs at the third codon position (figures 5.4.2 b&c), there are substantially more regions of the *Shigellae* genomes which show 'strong' deviations (defined as above) below the mean at the terminus and above the mean at the origin with a clear distinction of the two, the pattern 'flipping' from strong deviations below the mean to strong deviations above the mean about halfway between the origin and terminus. *S. sonnei* shows both a strong negative difference at the terminus and a strong positive difference from the mean at the origin. Interestingly EcC shows unshared strong negative deviations from the mean proportion of SNPs (figure 5.4.2c around Gene number 1000) which is in approximately the same location that it shows a strongly positive deviation from the mean SNP density – this region is showing a strong clustering of nonsynonymous SNPs.

		Codon Position		
		1st	2nd	3rd
Mean of *E.coli*	Min	10.2%	4.4%	71.5%
	Mean	14.3%	7.0%	78.8%
	Max	18.4%	10.4%	85.1%
Mean of *Shigellae*	Min	17.9%	11.7%	55.3%
	Mean	21.3%	15.4%	63.3%
	Max	24.9%	20%	69.9%

Table 5.4.2a – The minimum, maximum and mean proportions of SNPs observed at each codon position for both the mean of all *E. coli* and the mean of all *Shigellae*.

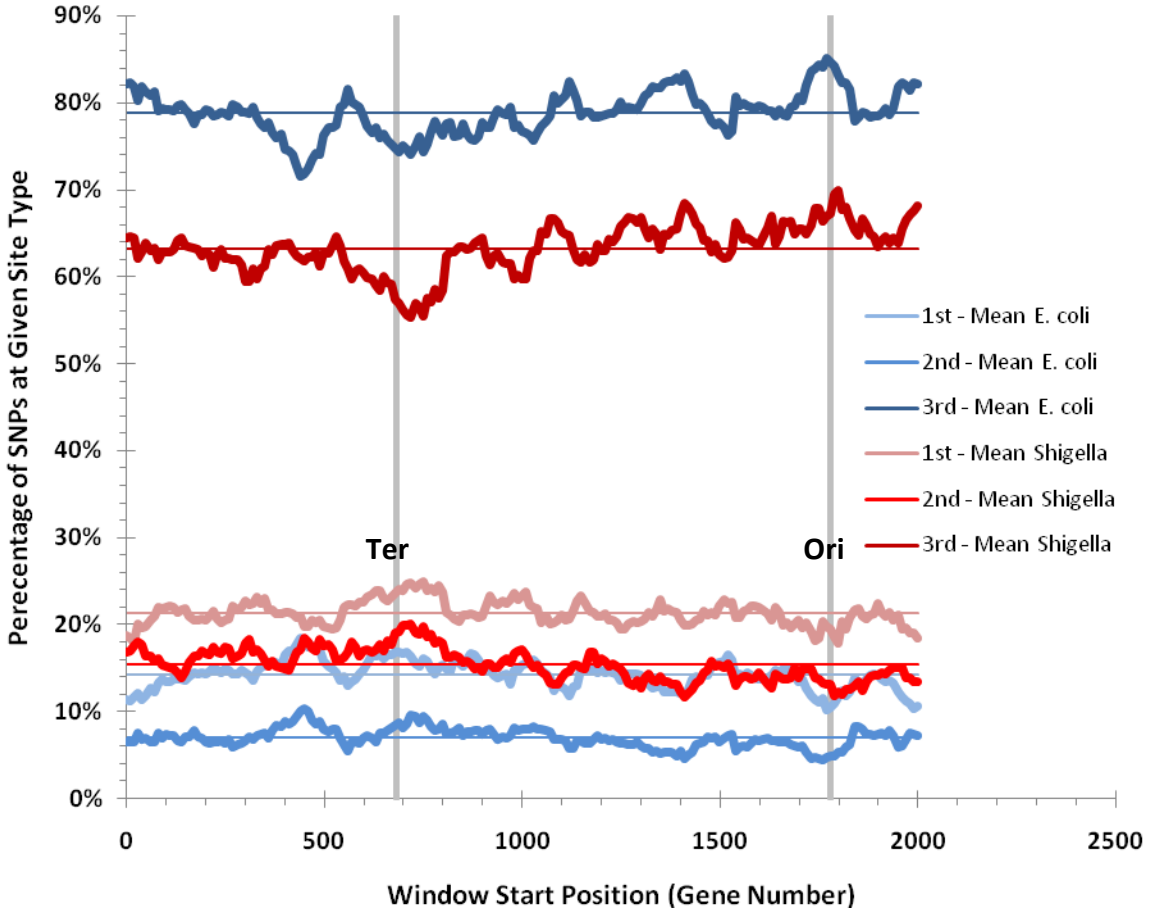


Figure 5.4.2a – Distribution of SNPs between codon positions along the aligned core orthologues, showing the mean of the *E. coli* (blue) and the mean of the *Shigellae* (red). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively). The fine horizontal lines reflect the overall mean proportion of SNPs at each codon position, for each group of taxa.

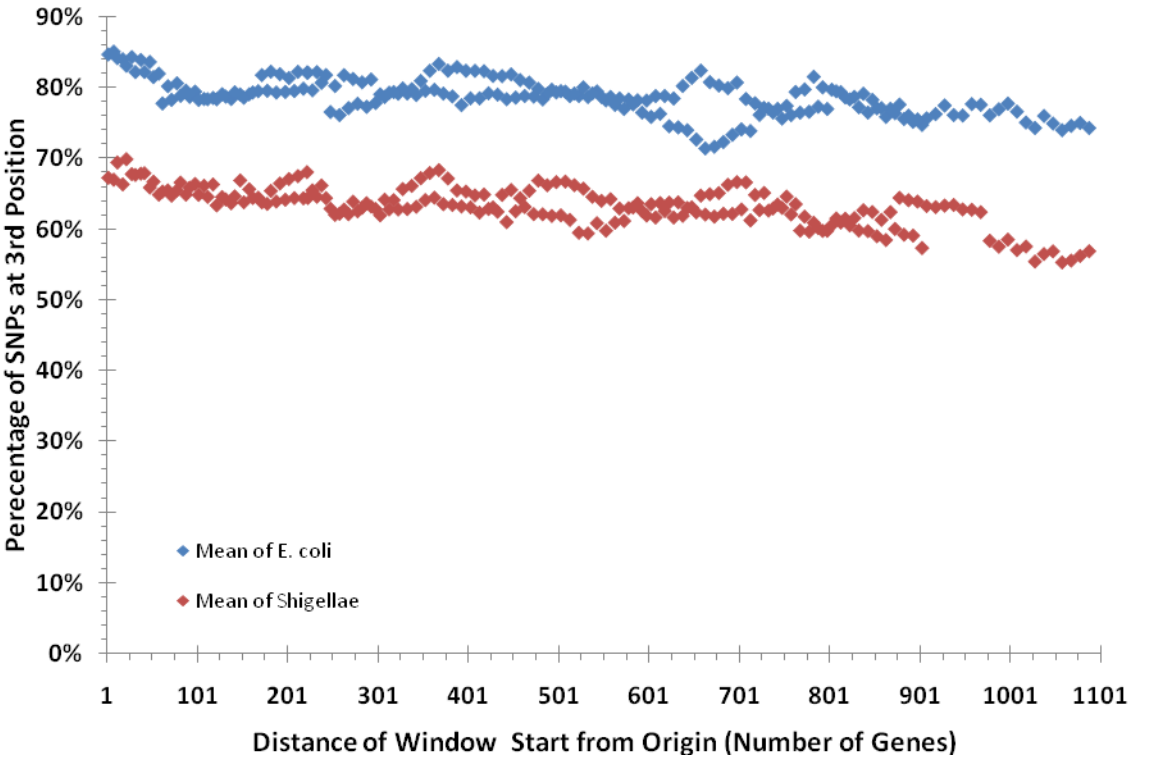
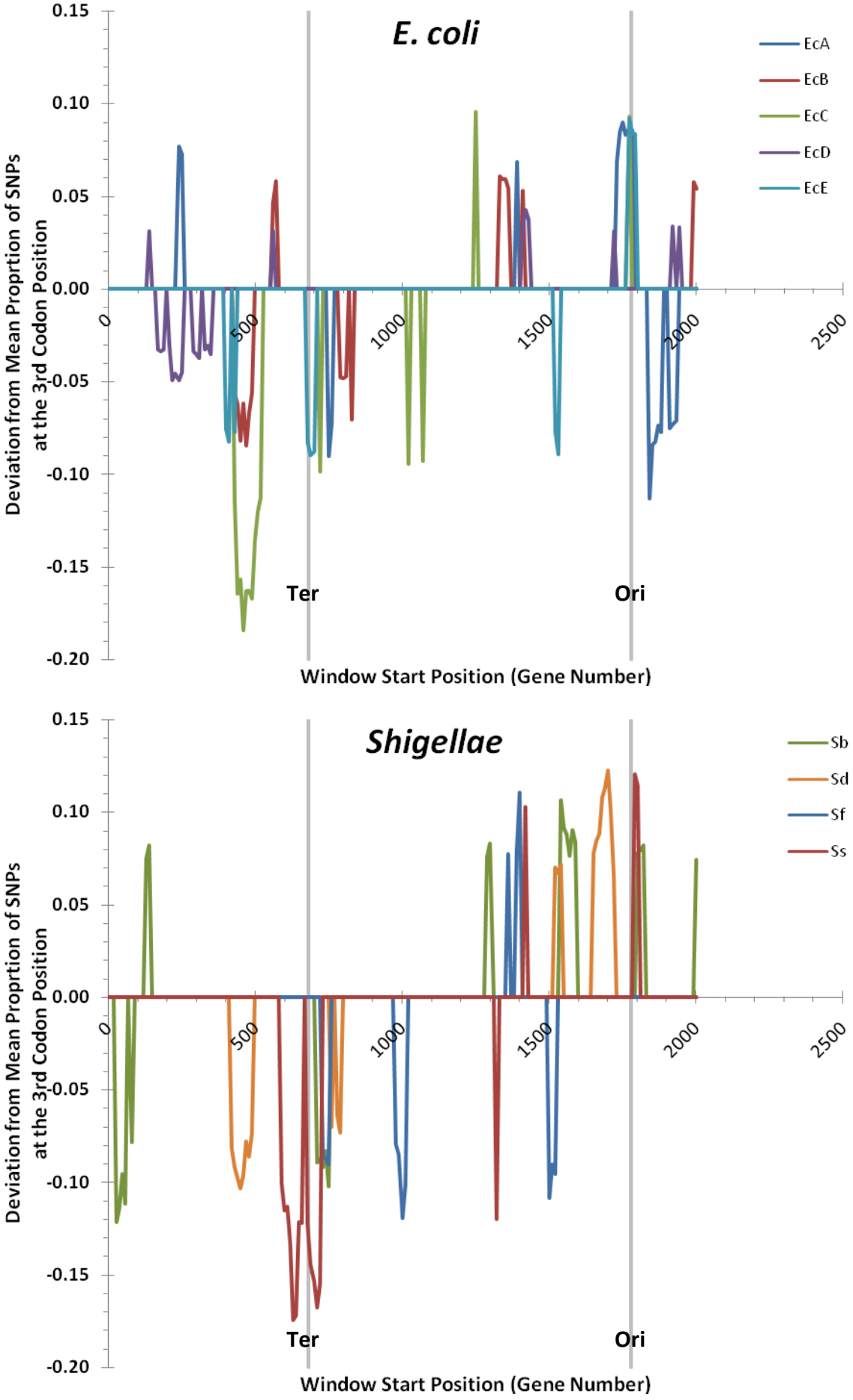


Figure 5.4.2b – Plot of the mean proportion of SNPs occurring at the 3rd Position for *E. coli* (blue) and *Shigellae* (red) against distance from the origin.



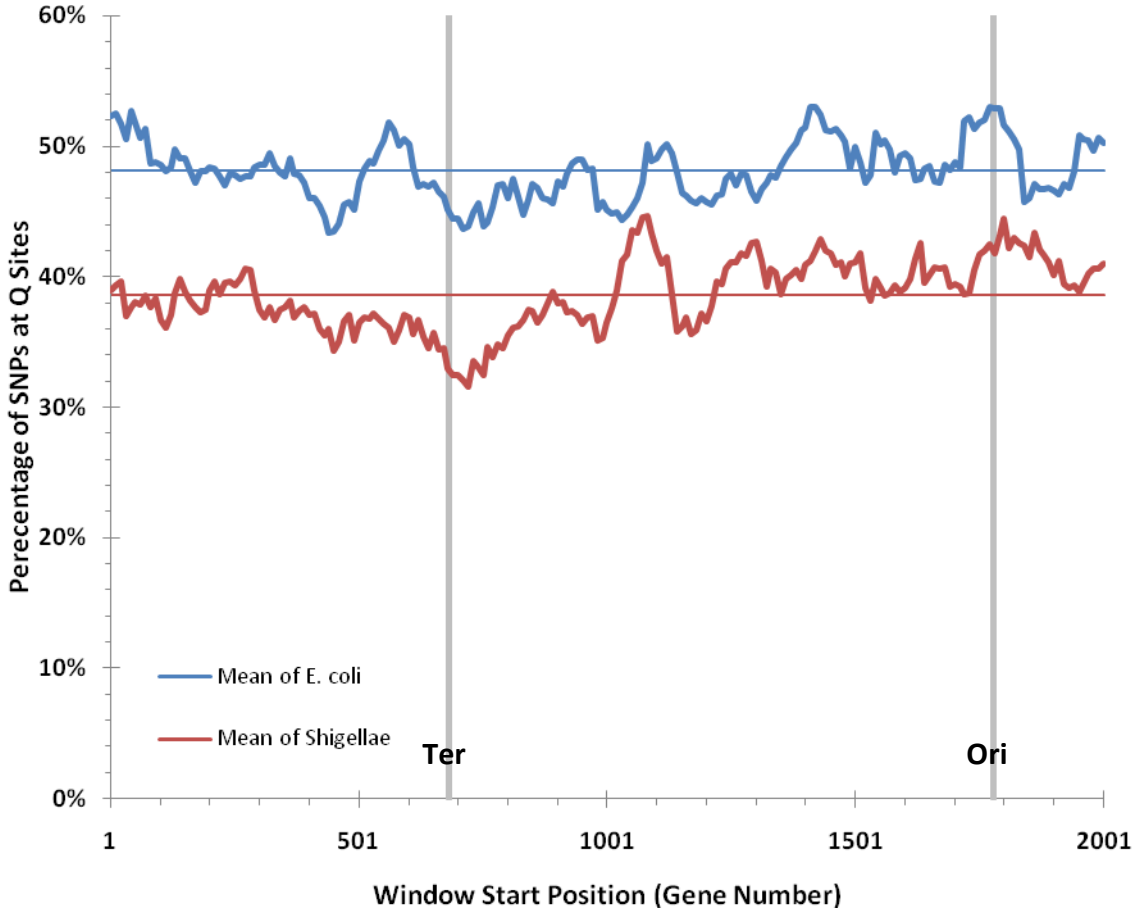
Figures 5.4.2 c&d – Plots of the 'Strong' deviations from the mean proportion of SNPs at third codon positions in both in each of the *E. coli* (a) and *Shigellae* (b). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).

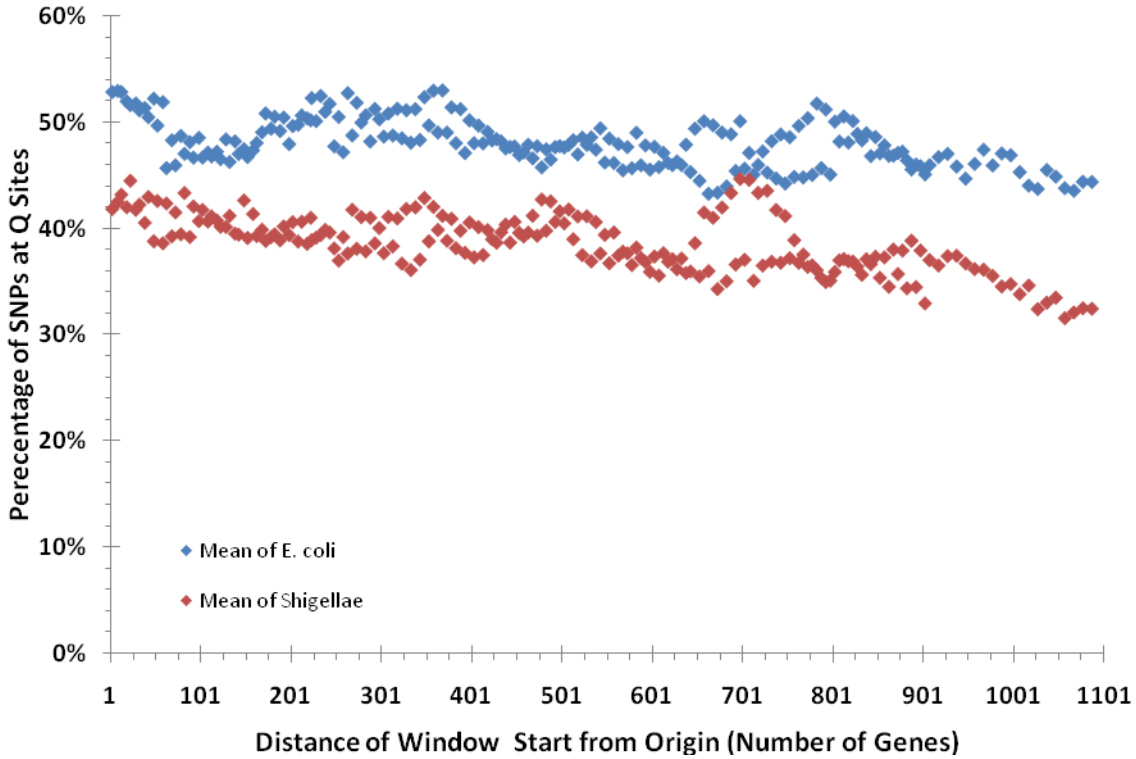
5.4.3 – Q Site Variation

As with the proportion of SNPs occurring at third sites, the *Shigellae* show a lower proportion of SNPs at the Q sites, indicative of fewer synonymous substitutions. Also evident from figures 5.4.3a & b (below) is the same propensity for fewer SNPs at Q sites near the terminus than observed at the origin, the trend being especially clear in figure 5.4.3b.

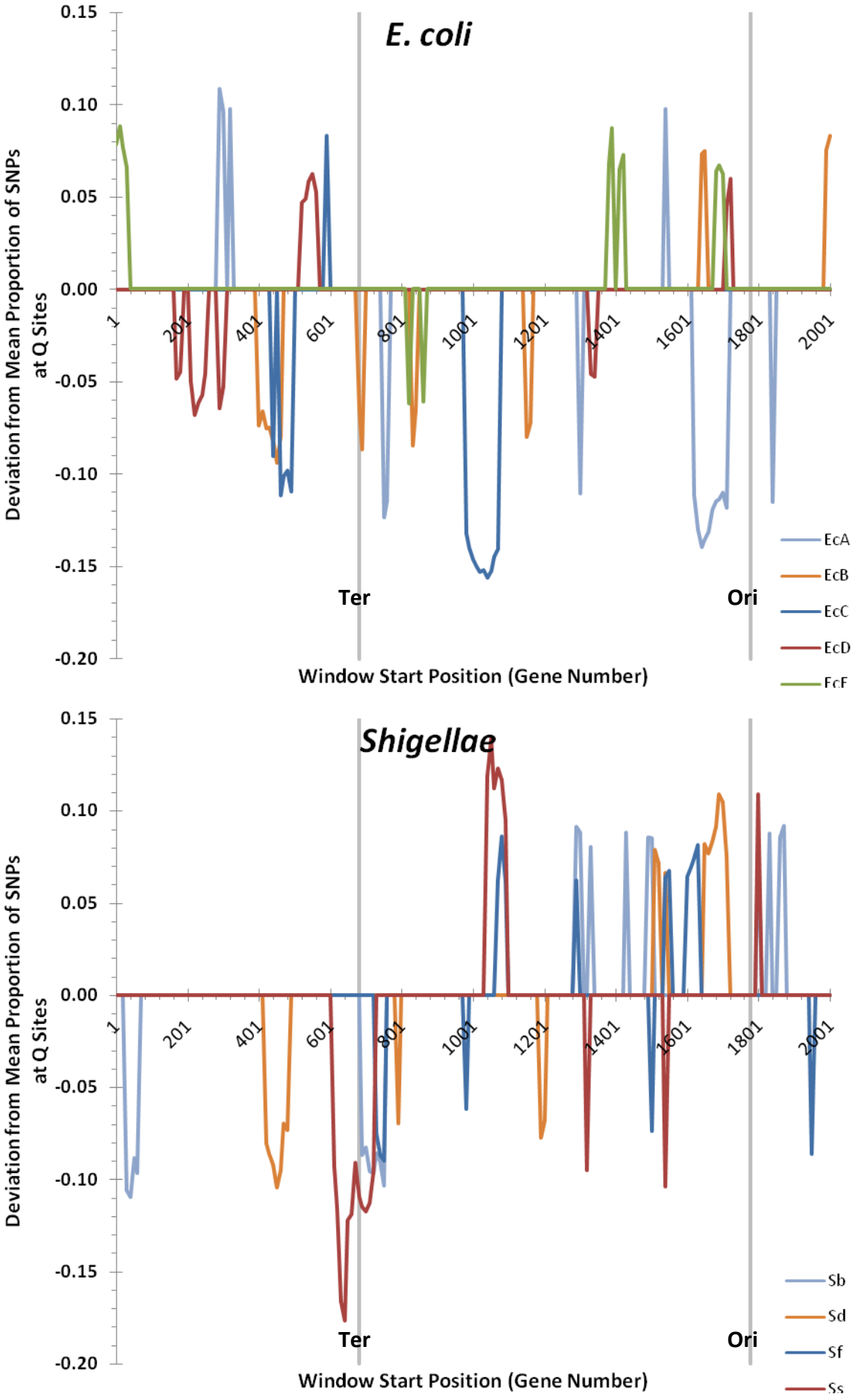
There is no clear bias between the origin and terminus within the *E. coli* in terms of the number of 'strong' deviations from mean proportion of Q sites, with many of the *E. coli* showing both strongly positive and negative deviations at both the origin and terminus, where such deviations are present. EcC for example shows both strongly positive and negative deviations only at the terminus, whilst the EcA shows the same but largely at the origin. The strongest deviation is negative between 990 and 1070, and represents a region in EcC with a proportion of SNPs at Q sites that is 15 percentage points lower, this region roughly agrees with the smaller region of low 3rd site SNPs and the region of high SNP density previously observed in EcC (figures 5.4.1b & 5.4.2b).

The *Shigellae* show strongly negative biases all along the genome, with the location and spread varying to some extent with species. *S. flexneri*, showing very few deviations, has negative deviations near both the origin and terminus. However in the *Shigellae* showing several strong deviations there is a trend towards positive deviations nearer the origin and negative deviations nearer the terminus.





Figures 5.4.3a & b – Plot of the mean proportion of SNPs occurring at Q sites for *E. coli* (blue) and *Shigellae* (red). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively) and the overall mean values for each group of taxa (horizontal lines) (a). And the same values plotted against distance from the origin (b).



Figures 5.4.3 c&d – Plots of the 'Strong' deviations from the mean proportion of SNPs at Q sites in both in each of the *E. coli* (a) and *Shigellae* (b). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).

5.5 – Variation in Metric Ratios

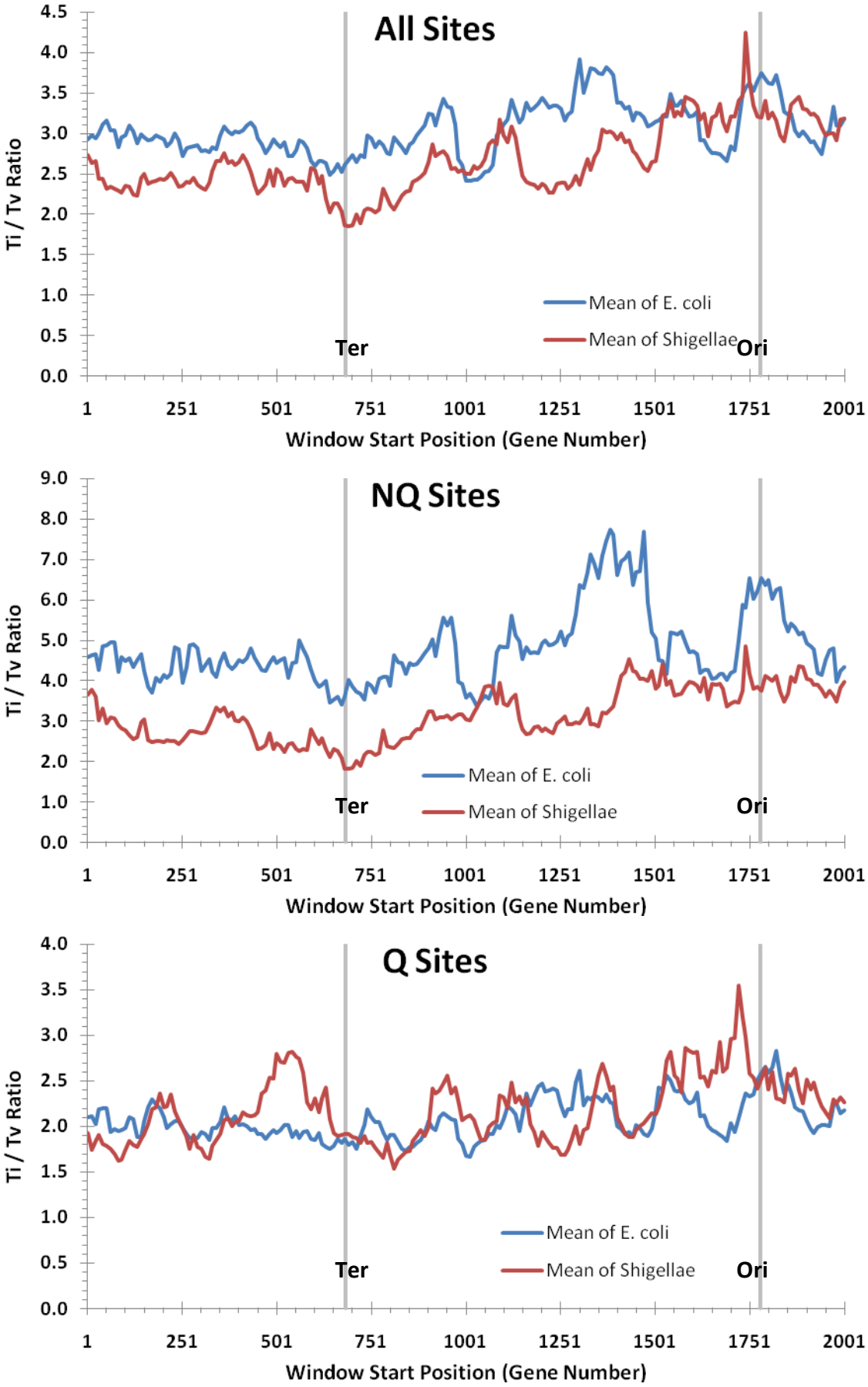
5.5.1 – Transition / Transversion Ratio

For much of the alignment the mean Ti/Tv ratio value for *E. coli* is higher than the ratio value for the *Shigellae*, reflecting the reduced purifying selection identified in the *Shigellae* (as compared to the *E. coli*) in Chapter 3. This difference however is almost completely absent in regions around the origin; given the relative abundance of DNA replication and other essential or highly expressed genes in this region this effect is likely a result of higher selective constraint at these loci.

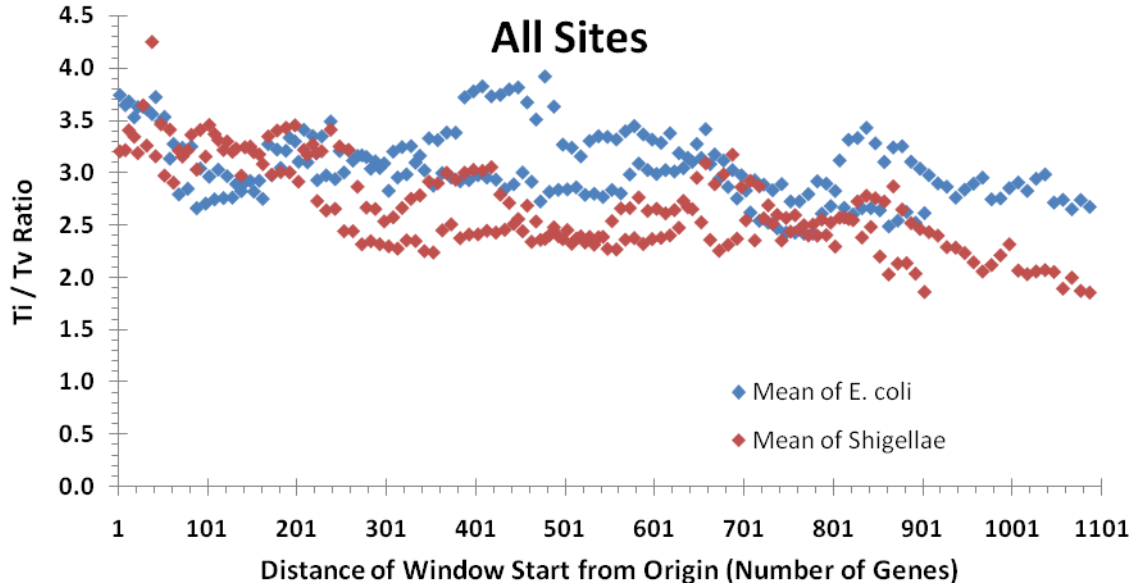
The differences in the values observed at NQ and Q sites, is again indicative of the differences observed earlier (Chapter 3). The NQ sites showing a much greater Ti/Tv ratio, reflecting the higher selective cost associated with the transversions at these sites. Along the alignment there is however a trend towards higher Ti/Tv values at the origin, a feature more pronounced in the *E. coli* than the *Shigellae*, again the inferred reduced selection in the *Shigellae* would manifest as a greater proportion of the more deleterious changes, in this case transversions – leading to a reduced Ti/Tv ratio. The region of markedly higher Ti/Tv ratio in the *E. coli* (~genes 1251-1501) is the result of several transversion types in this region with zero counts in several windows in each of the *E. coli*.

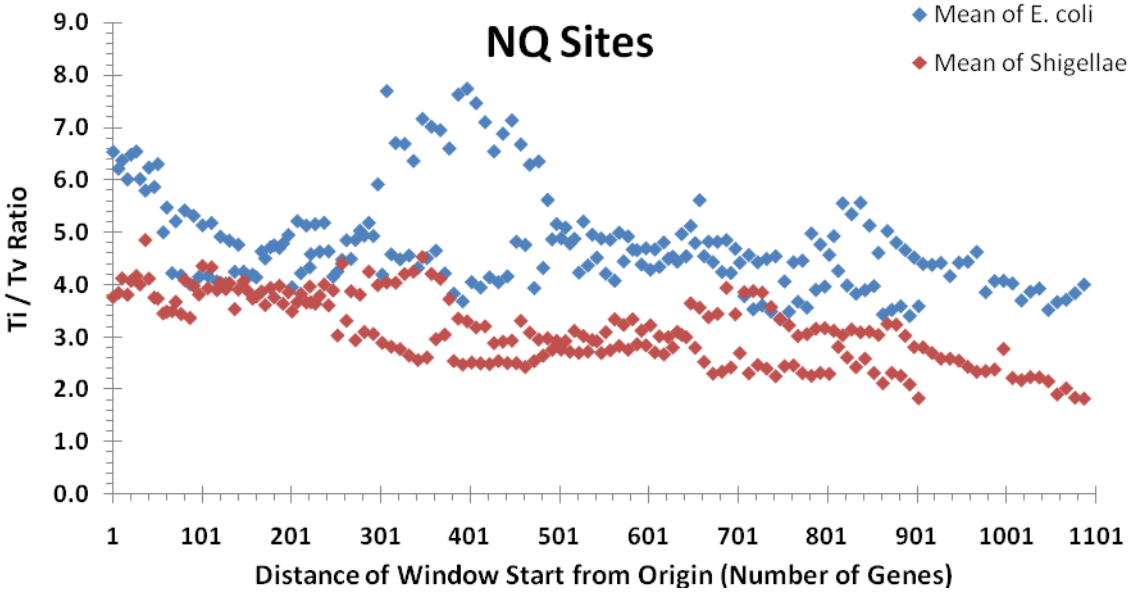
The pattern in the ratios at Q sites shows less of a distinction between the terminus and the origin, there is however a marginal increase in the ratio towards the origin. The *E. coli* show less fluctuation in the ratio value along the genome than do the *Shigellae*, which show two notable regions of increased Ti/Tv; one just before the terminus and one just before the origin.

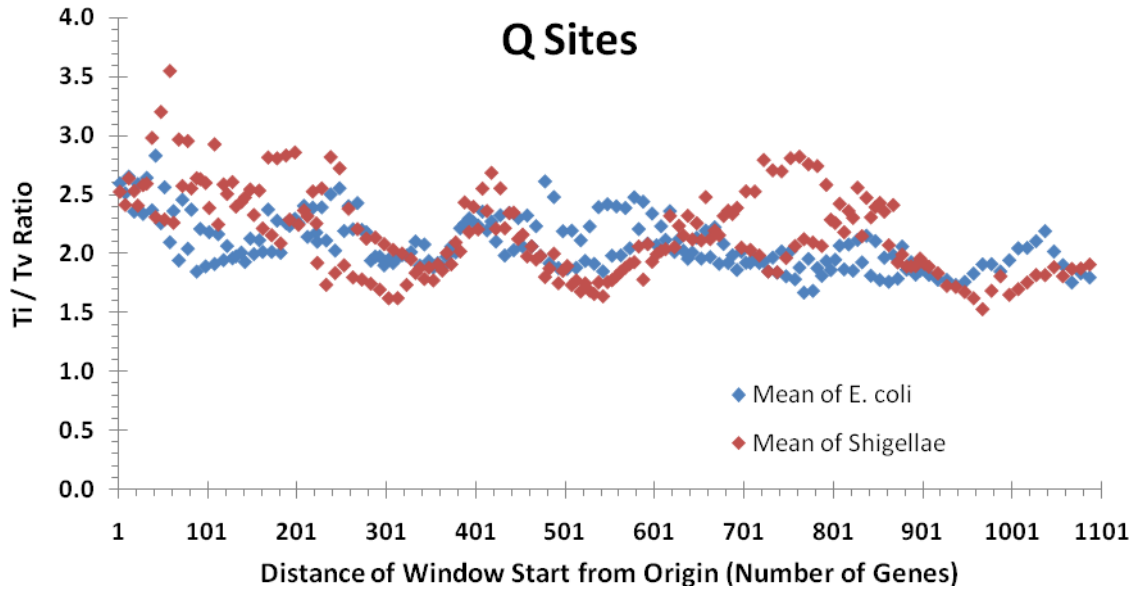
When examining the ratio plotted against distance from the origin, there is a clear pattern of higher proportion of transitions at the origin overall and at both site types, with NQ sites showing clear separation of the *E. coli* and *Shigellae* (higher ratio in *Shigellae*) and there being no discernable difference in the ratio at Q sites, reflecting the difference in selective constraint between the two sites and the resultant differences in the patterns of selection observed in the *Shigellae* and *E. coli.*



Figures 5.5.1 a, b&c – Plots of the Ti/Tv ratio for the Mean of *E.* coli (Blue) and Mean of the *Shigellae* (red). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).







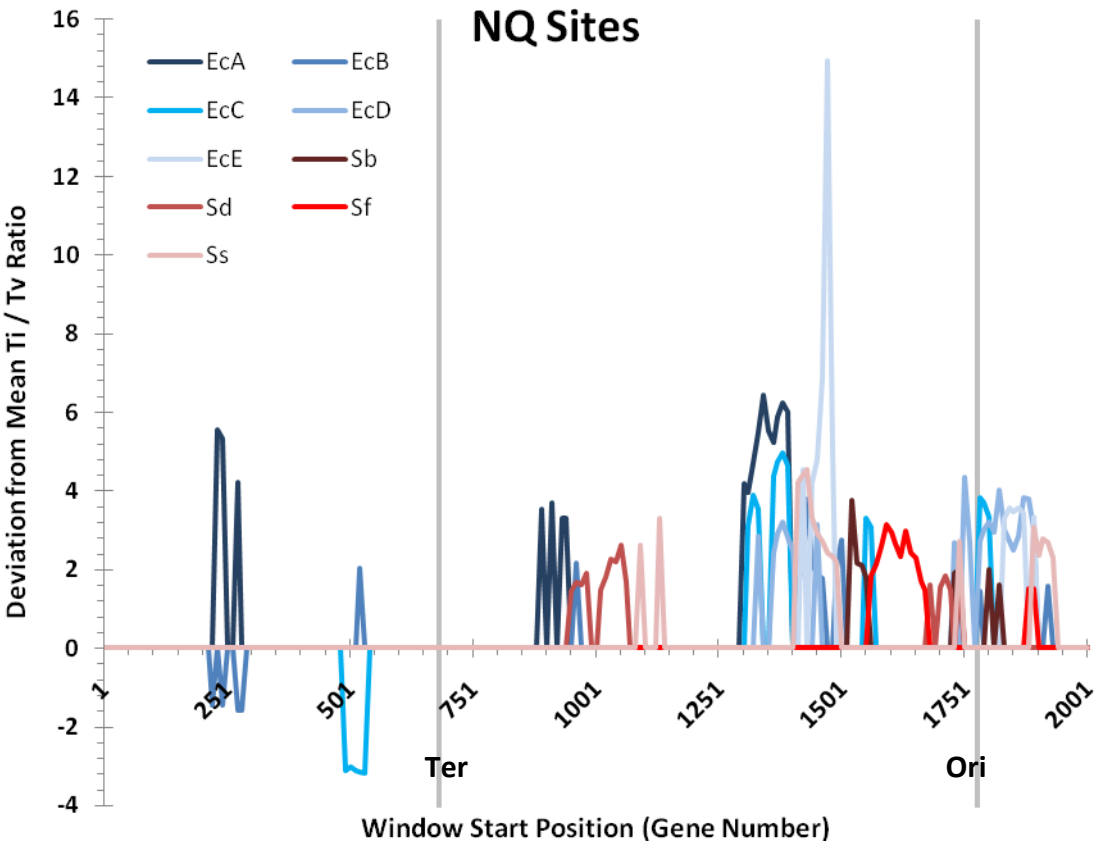
Figures 5.5.1 d, e&f – Plots of the Ti/Tv ratio for the Mean of *E.* coli (Blue) and Mean of the *Shigellae* (red) against distance from the origin.

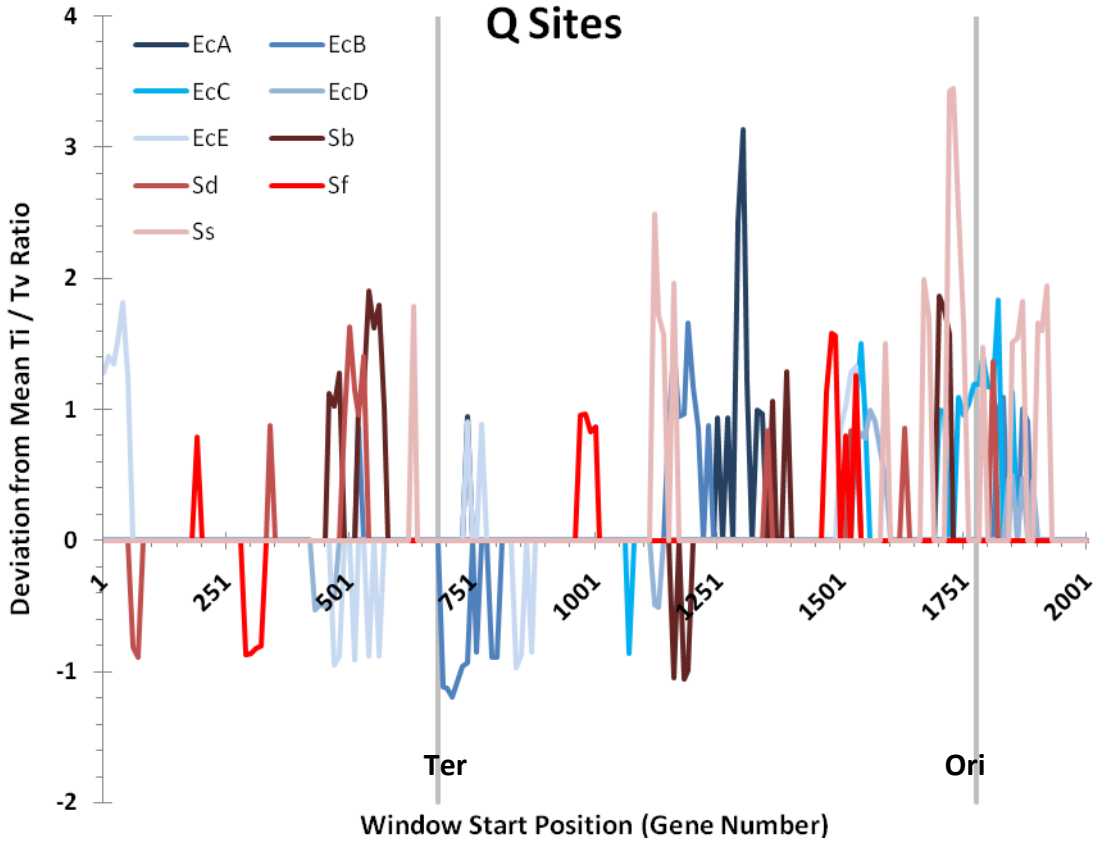
5.5.2 – Regions strongly deviating from the Mean Ti/Tv Ratio

At NQ sites there are relatively fewer *Shigellae* regions which deviate strongly, again defined here as the most extreme 10% of values (5% up and 5% down). *S. flexneri* shows only one region, greater than one sliding window, of strongly increased Ti/Tv ratio approximately from 1550 to 1700. These generally lower levels of deviation in the *Shigellae*, excepting *S. sonnei*, is likely a consequence of the reduced purifying selection experienced by the *Shigellae* resulting in the weakened bias observed in the mean of all *Shigellae*. *S. sonnei* shows more 'strong' deviations from the mean than do the other *Shigellae*, however the differences observed are neither as great nor as numerous as observed in the *E. coli*, possibly a signature of the less reduced purifying selection evident in *S. sonnei* due to its recent divergence and ability to survive in an environmental host. It is important to note that these patterns in the Ti/Tv ratio could also be explained by reduced horizontal gene transfer (HGT) due to reduced opportunity, *S. sonnei*'s ability to survive in an environmental host providing a slightly higher level of opportunity for HGT. However levels of HGT do not readily explain the origin centred biases seen above.

The *E. coli* at NQ sites show a pattern of greater and more numerous (the majority positive) deviations from their respective mean Ti/Tv ratios at and around the origin, consistent with the existence of essential and highly expressed genes closer to the origin which would be under greater purifying selection. A notable peak is that of EcE between 1411 and 1481, these windows show a deviation from the mean of +5 to +15, however such an extreme bias may likely be an artefact of the method given that some of these windows have zero counts for between one and five of the eight transversion types. The most extreme window at 14, showing a Ti/Tv ratio of 20, has a zero count for five of the transversion types, severely biasing the ratio towards transitions.

At Q sites the strong regions in the *Shigellae* are scattered along the genome, there is a slight bias towards the origin but there are also strongly higher regions at the terminus. The *E. coli* show strongly deviating windows with a bias towards positive deviation at the origin and negative deviation at the terminus.





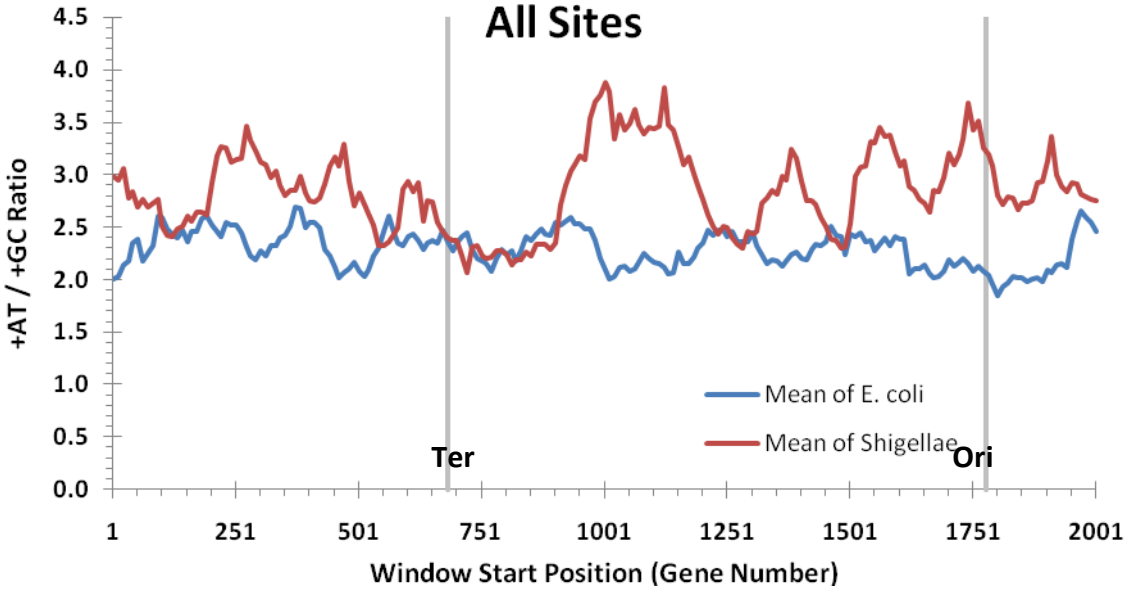
Figures 5.5.2 a & b – Plots of the 'Strong' deviations from the mean Ti/Tv Ratio at NQ and Q sites. Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).

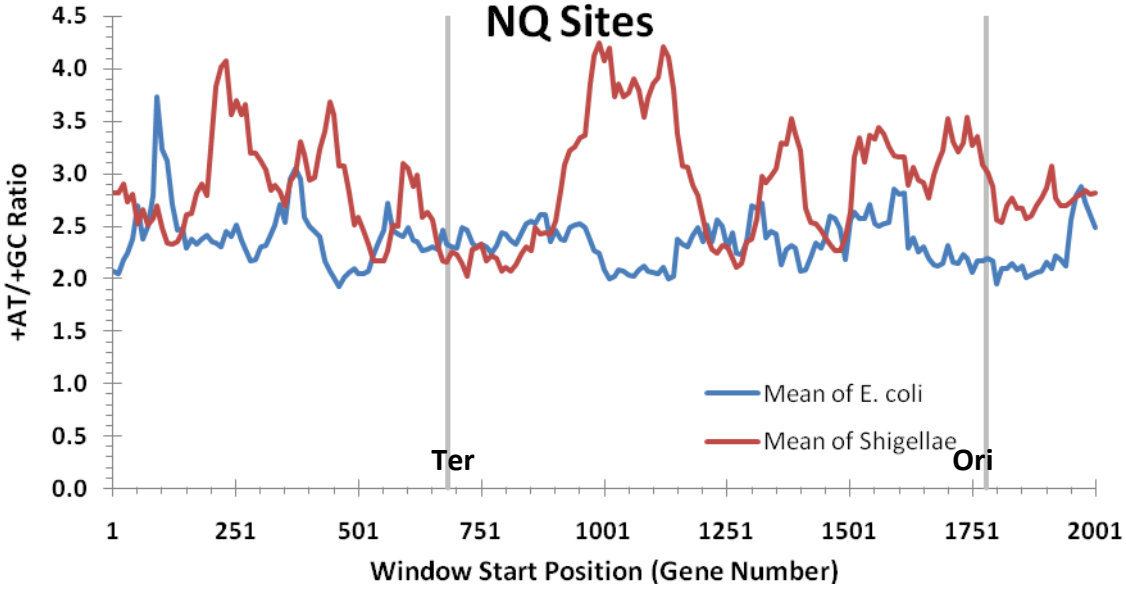
5.5.3 – Ratio of AT to GC enriching SNPs

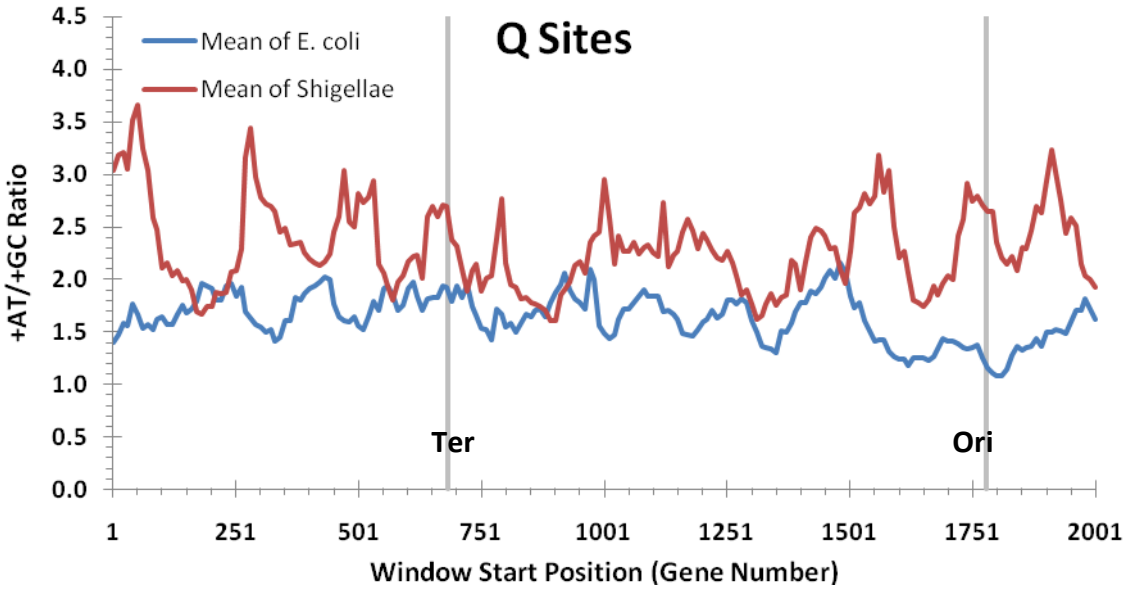
At both all sites and NQ sites the *E. coli* show less variation in the +AT/+GC ratio than do the *Shigellae*, in line with the effects associated with the reduced purifying selection in the latter, in addition all points along the genomes show ratio values consistent with AT enrichment. The *Shigellae* also show a large region of increased +AT/+GC ratio between gene numbers 850 and 1250, which is evident at NQ sites, but not Q sites. However this likely reflects the generally higher levels of variation in the +AT/+GC ratio at NQ and 'all sites' in the *Shigellae* and the higher mean value of the ratio, as this region shows no zero-counts of any SNP type in these windows. Both the *E. coli* and the *Shigellae* show no strong differentiation between the origin and the terminus in this metric at all sites or at NQ sites (figures 5.5.3d & e).

At Q sites, the *E. coli* show a +AT/+GC ratio approaching parity at the origin, which is in line with the known trend for the location of essential and highly expressed genes and the previously mentioned GC richness of the preferred codons in *E. coli* and *Shigellae*. Mutations in essential or highly expressed genes would suffer from inefficiencies in translation as result of the use of unfavoured AT rich codons and therefore be under greater purifying and show a preferential purging of AT enriching SNPs in order to maintain the status quo. As at NQ sites the *Shigellae* show much greater variation and a generally higher level of AT enrichment than the *E. coli* but no apparent bias related to the locations of the origin or terminus. The *E. coli* mean +AT/+GC ratio significantly and strongly correlates with genome position ($r = 0.521$ & $p < 0.001$) and whilst the *Shigellae* mean +AT/+GC ratio correlates significantly, the correlation is not strong ($r = -0.225$, $p = 0.001$).

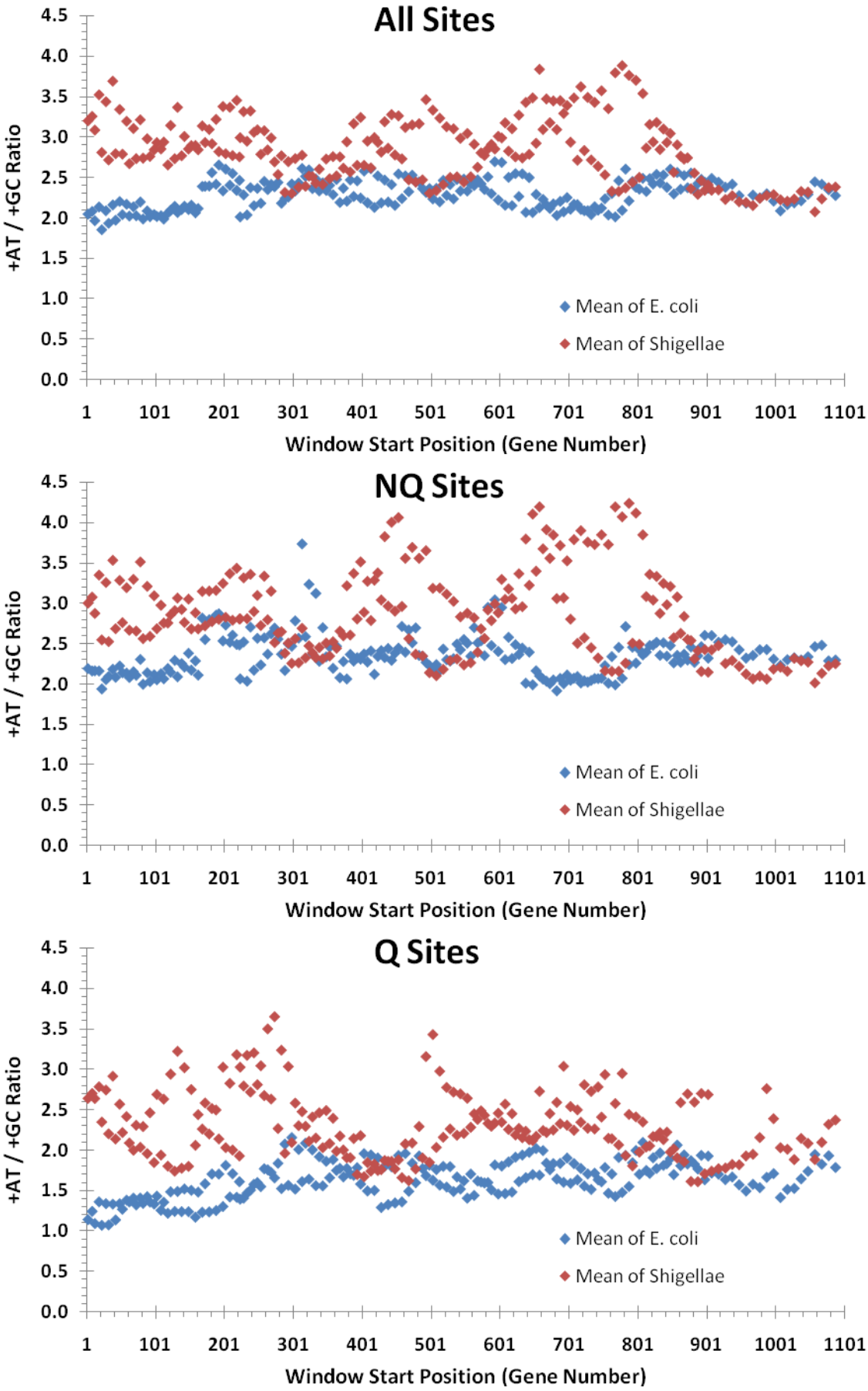
Overall the lack of strong location bias in the +AT/+GC ratio, evident in figures 5.5.3d, e & f, likely reflects selection for more global genome features associated with AT content rather than for specific features relating to the maintenance of coding sequences.







Figures 5.5.3 a, b & c – Plots of the +AT/+GC ratio for the Mean of *E.* coli (Blue) and Mean of the *Shigellae* (red). Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).

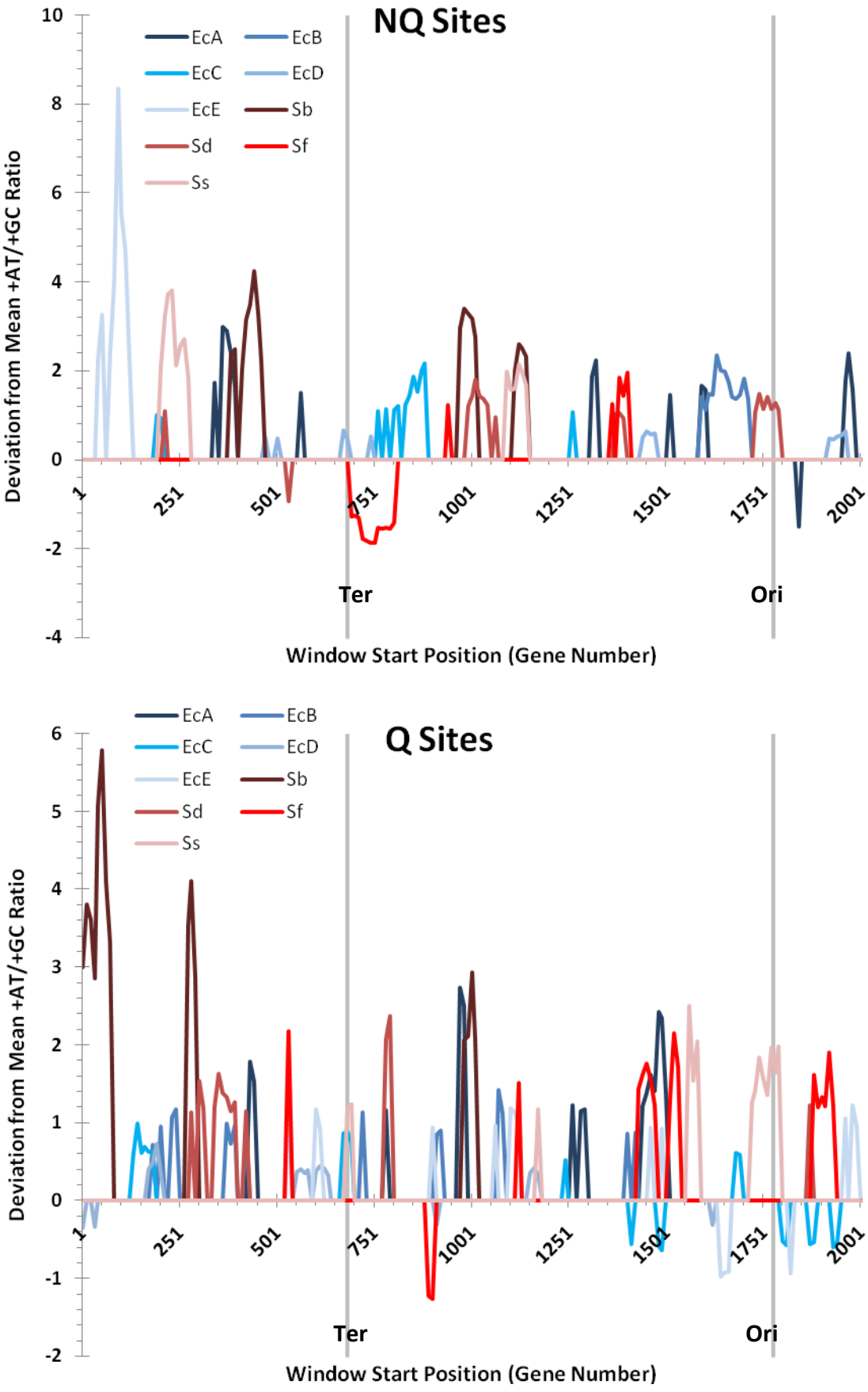


Figures 5.5.3 d, e & f – Plots of the +AT/+GC ratio for the Mean of *E.* coli (Blue) and Mean of the *Shigellae* (red) against distance from the origin.

5.5.4 – Regions strongly deviating from the Mean +AT/+GC Ratio

At both NQ and Q sites there is no clear distinction of the Origin or Terminus in terms of the direction of abundance of deviations of the +AT/+GC ratio from the mean (figures 5.5.4a & b), this is not unexpected given the lack of trend observed in the +AT/+GC ratio versus distance from the Origin (figures 5.5.3d, e & f). The high levels of variation in the ratio in the *Shigellae* are reflected by the generally higher magnitude of their strong deviations versus the *E. coli*, this likely reflects the previously identified weaker purifying selection in the *Shigellae*, the lower selective constraint permitting greater variation in the +AT/+GC ratio along the genome.

There are two outstanding regions; the first, at NQ sites, is between gene 41 and 131 in EcE, this region shows a positive deviation from the mean +AT/+GC ratio of up to 8.35. Examination of the SNP counts reveals that this whilst this region has high counts of C→T and G→A, these are not obviously higher than adjacent regions, however it does have much lower values of T→C and A→G than neighbouring windows. The second, at Q sites, is between gene 1 and 71 in *S. boydii* (Sb), as with EcE this region shows no higher C→T and G→A values than it's neighbours, however in this case there is a notable paucity of A→N SNPs.



Figures 5.5.4 a & b – Plots of the 'Strong' deviations from the mean +AT/+GC Ratio at NQ and Q sites. Showing the approximate location of the origin and terminus (at Gene Numbers 1779 and 682, respectively).

5.6 – Anomalous Region in EcC

In both the SNP density and the SNP site distributions there is a clear region showing strong deviations from the mean in EcC, the location of this region is detailed below (Table 5.6a) for each of the metrics where it is observed. Both the SNP type ratios do not show a strong deviation from the mean in this region however, as can be seen below (Figure 5.6a) the Ti/Tv ratio shows a large negative deviation from the mean between genes 980 and 1070. The +AT/+GC ratio also shows a negative deviation in this region, which closely matches the location of the Ti/Tv deviation, however it is part of a much wider deviation stretching from gene 980 to gene 1150.

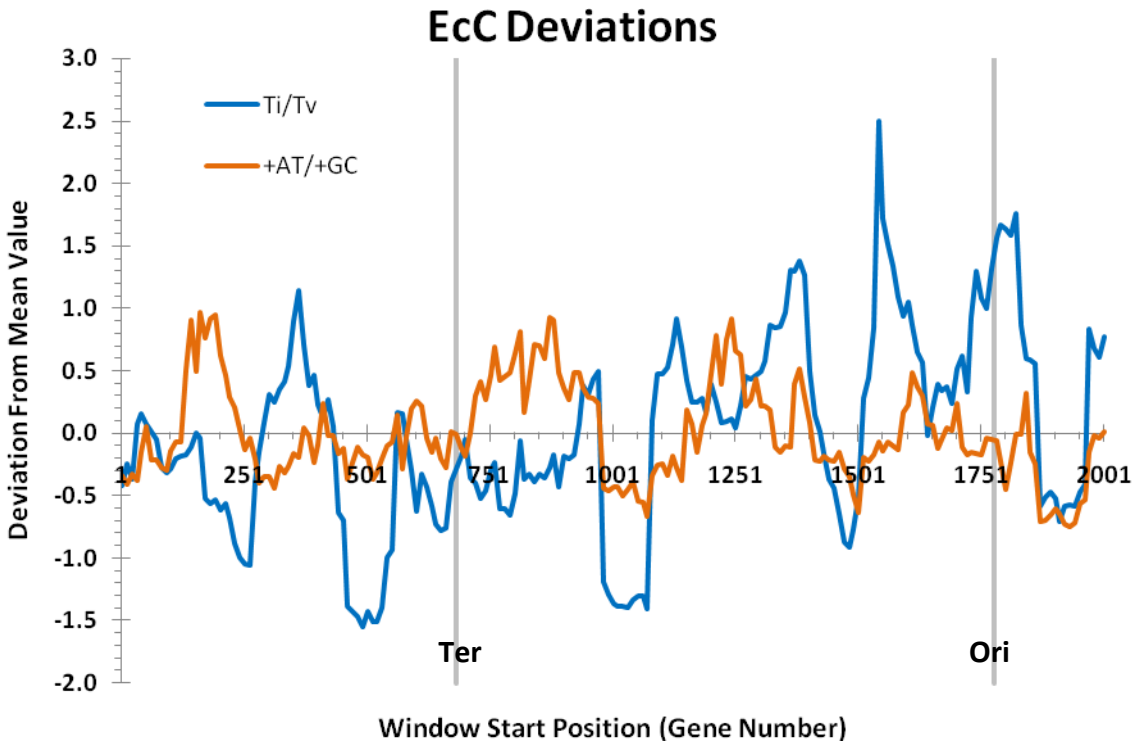


Figure 5.6a – Plot of the deviations from the mean value for the Ti/Tv and +AT/+GC ratios in EcC

The phylogenetic tree (figure 5.6b) of the deviated region shows no major difference in the positioning of the EcC taxon from the 2098 orthologue tree generated in Chapter 3, there is a polytomy towards the base of the tree that is resolved in the Chapter 3 tree, however that has no bearing on the origins of the rapidly evolving region in EcC.

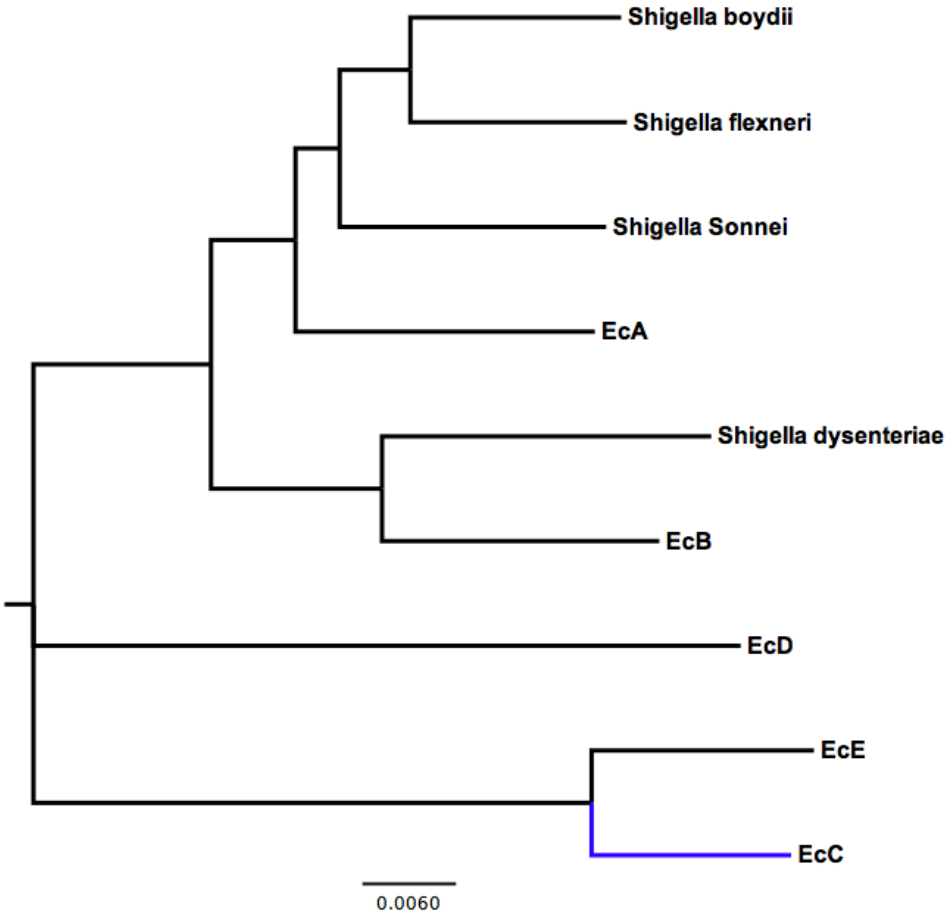


Figure 5.6b – Neighbour Joining Tree of Genes 991-1031 (the core region covered by all the deviated metric ratios). The branch to EcC is highlighted in Blue.

Metric	Region Location	Mean Region Value	Mean Flanking Region Values	Taxon Mean
SNP Density (SNPs/Kbp)	1.03 – 1.12Mbp	5.297	2.837 / 2.436	2.704
SNPs at 3rd Sites	Genes 980 – 1070	68.2%	71.7% / 80.3%	76.9%
SNPs at Q Sites	Genes 990 – 1070	32.4%	46.2% / 49.2%	47.0%
Ti/Tv	Genes 980 – 1070	1.704	3.096 / 3.527	3.046
+AT/+GC	Genes 980 – 1150	1.893	2.774 / 2.608	2.266

Table 5.6a – The mean values of several metrics of selection within and flanking the region identified throughout the analysis. In each case the flanking region is defined as a region of the same size both immediately up and down-stream.

Table 5.6a shows clearly that there is a very distinct signature of increased evolution in the region identified, there being a roughly twofold increase in SNP density accompanied by an increase in the proportion of nonsynonymous SNPs, as indicated by the reduced proportion of SNPs at the 3rd codon position or Q sites. Additionally these SNPs are also more likely to be transversions than SNPs in the flanking regions. The region showing a reduced +AT/+GC ratio is approximately twice the size of the region showing a deviation in all the other metrics.

The majority of genes identified in this region by BLAST search are involved in central and intermediary metabolism (33 of 56) of which 15 are components of NADH dehydrogenase I. The remainder of the genes are 10 hypothetical genes, 9 outer membrane genes (including one antibiotic resistance membrane transport protein), 3 Nucleotide metabolism genes and 1 regulatory protein.

5.7 – Summary of Results

- The Synteny of the genomes is largely conserved, with major rearrangements and inversions being highly symmetric and centred on either the Origin or Terminus of replication.

- The nucleotide composition of the core genome shows a clear bias towards decreased AT (increased GC) richness at the Origin, which falls steadily with the minimal bias being approximately centred on the Terminus.

- The density of the SNPs along the genome varies considerably, with several genomes having regions that show a notably higher or lower density than the overall genome mean. However there is no obvious clustering of regions showing notably higher or lower SNP density that is centred on either the Origin or Terminus.

- The abundance of SNPs at the more synonymous 3rd codon position and Quartet (Q) nucleotide sites is higher around the Origin than the Terminus.

- A greater proportion of SNPs around the Origin are Transitions (Ti) than closer to the Terminus, with the *E. coli* showing a consistently higher proportion of Transitions than the *Shigellae*.

- The bias between AT enriching and GC enriching SNPs varies along the genome, and shows a slight correlation to distance from the origin, showing a reduced enrichment bias within 300-400 genes either side of the origin. There is considerably greater fluctuation in the +AT/+GC ratio along the *Shigellae* genomes than there is along the *E. coli* genomes.

- EcC shows a region with a significantly higher than average SNP density within which a greater proportion of the SNPs are nonsynonymous, also showing a greater bias towards transversions and reduced AT enrichment compared to flanking regions.

5.8 - Discussion

5.8.1 – Synteny and Nucleotide Composition

The existance of relatively large inversions in the *Shigellae* genomes is unsurprising, the reduced purifying selection previously documented in organisms with reduced effective population size (Woolfit and Bromham 2003; Woolfit and Bromham 2005) and the effects observed (Chapters 3 & 4) in the *Shigellae* would mean that such rearrangements are less likely to be purged, especially given the lack of opportunities for individual *Shigellae* to recombine with anything other than a close relative due to their intracellular lifestyle. Additionally the highly symmetric nature of the rearrangements, which are centred on either the Terminus or the Origin, results in no net change in the position of a gene in the genome relative to either of these points. The majority, and the largest, are located at the terminus, which of the two symmetric points of the genome contains fewer 'essential' genes or genes involved in genome replication (Couturier and Rocha 2006). Indeed it has been noted that the majority of ribsomal protein genes are located in the first third of the genome, from the origin, in both directions (Rocha 2004). A bias of ~70% around the origin to ~5% around the terminus was noted in an analysis of 68 bacterial genomes.

These inversions have had no observable effect upon the nucleotide content of the genome, which shows a strong and consistent bias towards GC richness at the origin, which in turn is likely the result of the bias towards GC rich codons in *E. coli* – the more highly expressed genes around the origin showing a greater bias towards 'preferred' codons, improving translational efficiency (Kurland 1991). The lack of any observable difference between the patterns observed in the *E. coli* and *Shigellae* is unsuprising given their short divergence times (10 – 270 Kya) and the consequently small differences in their extant genome composition.

5.8.2 – SNP Density & Position

There is considerable variation in the SNP density within each taxon and between the taxa, with mean SNP densities ranging from 1.73 SNPs/Kbp (Ss) to 6.49 SNPs/Kbp (EcD). The variation between the taxa is attributable to the different branch lengths

associated with each taxon and therefore the amount of evoltionary time that is being examined in each case, the variation within a given taxon reflects a combination of the largely random location of incurred mutations and variations in selective pressures associated with different genes or, as is likely given the window size, groups of genes. There is also no consistent pattern deviation from the mean SNP density with repect to the distance from the origin, the absence of lower SNP densities in more essential or highly expressed genes does not have an obvious explanation, there may either be insufficient SNPs to resolve any differences or the differences may only be observable over longer evolutionary distances.

There is little or no separation of the *Shigellae* and *E. coli* in terms of the mean deviation in SNP density, however across the whole core genome the *Shigellae* show a marginally higher mean SNP density than the *E. coli* (2.67 & 2.47 respectively), whilst this is indicative of the previously observed reduced effective purifying selection in the *Shigellae* the difference is not significant.

There are three notable taxa in terms of SNP density; EcD which shows two regions with below average SNP density, EcA which shows a region of high SNP density close to the origin and EcC which shows a region of high SNP density around 1Mbp along the alignment. It is not immediately apparent why there should be a region of extremely high SNP density in EcA or regions of extremely low SNP density in EcD, however the region of high SNP denisty in EcC is discussed later.

The SNP site distribution shows a clear bias with respect to distance from the origin, there being a greater proportion of SNPs at the more degenerate 3rd positions and Q sites around the origin with the proportion at these sites gradually falling to a minimum around the terminus. This pattern reflects the greater level of purifying selection that these genes are likely to be under, as such SNPs at 1st & 2nd positions or NQ sites are more rapidly purged around the origin than at the terminus. This only explains part of the pattern however as an effect based solely upon features associated with the origin should show a patten more tightly associated with the origin and the terminus should show little

differentiation from the rest of the non-origin genome. It has been shown that there are features of the genome which confer similar terminus centred selective biases on the genome – the clustering of binding sites for chromatin structure regulation (Ussery, Larsen et al. 2001) and the terminus centred skew of GC3 (Daubin and Perriere 2003) both of which are linked to the correct function of replication termination. This suggests that a combination of factors is responsible for the patterns observed however the consistency of the trend along the core genome suggests that either the factors involved are themselves linked or that a mechanistic factor may be involved.

The *Shigellae* show a lower proportion of SNPs at both 3rd positions and Q sites along the alignment whilst retaining the distinction between the origin and terminus, this suggests that the *Shigellae* are experiencing less efficient purifying selection than the *E. coli*, which is compatible with the observations in Chapters 3 & 4.

5.8.3 – Metric Ratios

The Ti/Tv ratio shows a clear distinction of the Origin and Terminus at NQ sites with there being a greater bias towards transitions at the origin, reflecting greater purifying selection acting on the more highly expressed or more essential genes. Unexpectedly, given the degeneracy of Q sites, there is a slight trend with distance from the origin in the Ti/Tv ratio, with a transition bias at the origin. It is unclear why there is such as bias, given that the selective costs of both transitions and transversions are, in terms of amino acid encoding, selectively equal at Q sites.

There is a clear distinction between the *Shigellae* and *E. coli* along the whole alignment at NQ sites, with the *Shigellae* showing a lower Ti/Tv ratio reflecting the previously identified reduced purifying selection (Chapter 3). However at Q sites, were there is little or no selective distinction between transitions and transversions there is no observable difference between the *Shigellae* and *E. coli* again mirroring, along the whole alignment, the patterns observed in Chapter 3.

The *E. coli* show a strong and significant correlation of the +AT/+GC ratio with distance from the origin at Q sites but not at NQ sites, the correlation for the *Shigellae* with distance

from the origin being nonsignficant (NQ sites) or weak (Q sites), it is unclear why this distinction is only significant at Q sites, as it would be expected, based upon the Chapter 3 results, that the distinction be evident at both site types.

The *Shigellae* are distinguished from the *E. coli* by a generally higher and more variable ratio value, consistent with reduced purifying selection resulting in accumulation of AT enriching SNPs. Additionally the *Shigellae* show a greater number of 'strong' devations from their mean ratio value than the *E. coli*, reflecting the greater variation in the +AT/+GC ratio.

5.8.4 – Anomalous EcC Region

The increased density of SNPs and the greater proportion of nonsynonymous changes observed, in the region from approximately gene 980 to gene 1070, is matched by shifts in the metric ratios Ti/Tv and to a lesser extent +AT/+GC. Ti/Tv shows a marked shift towards a greater proportion of transversions in this region than in either flanking region or the overall mean value and +AT/+GC shows a shift towards a greater proportion of SNPs being GC enriching.

Aside from the +AT/+GC ratio, all the other metrics portray a picture of increased sequence evolution in this region, which can potentially be explained by different mechanisms; this region of EcC might have undergone Lateral Gene Transfer from EcE and the new genetic information is sufficiently different from EcC that there is a strong selective pressure to restore the sequence features to the average or 'normal' for EcC, alternatively this region has not been acquired from another organism but is evolving rapidly as it is either undergoing genetic drift or it is under diversifying or adaptive selection.

Given the absence of any notable change in the topology of the phylogenetic tree based on this region from the overall alignment tree, it is unlikely that EcC has acquired this region from outside of the Uropathogenic *E. coli* (UPEC) of which EcE is also a member, and so amelioration of genetic differences from the genome 'norm' cannot explain the increased sequence evolution.

This distinction between increased genetic drift and adaptive evolution is somewhat more troublesome, however the +AT/+GC ratio provides an insight as it shows patterns of evolution more consistent with increased purifying selection relative to the mean. This better fits the adaptive evolution scenario; under adaptive evolution it would be expected that there would still be some level of selective constraint upon the changes. The +AT/+GC ratio correlates strongly with the mean metabolic cost per amino acid change (Chapter 4) whilst Ti/Tv does not, this suggests a scenario whereby there is adaptive evolution of the genes in this region with a selective constraint upon the metabolic cost of the amino acid changes.

This constraint on the metabolic cost of the amino acids used makes sense in light of the functional categories of the genes identified in this region, approximately a quarter of which are components of the NADH dehydrogenase complex, an electron transport protein in the respiratory chain of bacteria (Dancey and Shapiro 1976), which is likely to be under high selective constraint to minimise the metabolic cost of any amino acid changes. Additionally the 9 outer membrane genes observed to be in the region may be the source of the adaptive signal as outer membrane proteins are likely antigens for recognition by the host immune system.

5.8.6 – Overall Conclusions

In general the patterns of evolution observed along the alignment have supported the patterns identified in Chapters 3 & 4, and demonstrated that the differences between the *Shigellae* and *E. coli* are not confined to a subsection of the genome but apply along the whole alignment. Additionally, where there is a strong separation of the *Shigellae* and *E. coli* there is a clearer distinction between the values for the metric or measurement at the origin and terminus, providing support for the distinction of these two regions being due to increased purifying selection around the origin as a result of the clustering of highly expressed and essential genes.

These results are also supported by an analysis by Touchon et al (2009) wherein they observe a dearth of recombination events towards the terminus as well as a reduction in

GC content. The latter is reflected by the increase in +AT/+GC ratio observed with distance from the origin, the former is consistent with both +AT/+GC and Ti/Tv results as the lack of recombination events would result in a reduced ability to bring in potentially pre-selected mutations from the donor species or strain and therefore a reduced ability to rescue the more deleterious nucleotide changes.

It is important to note that the effects on the pattern of nucleotide changes observed along the genome could potentially be, in part, explained by mechanistic effects. Although it is unlikely to result in the systematic differences observed, it has been suggested (Rocha 2004) that the chronological separation of the replication of the origin and terminus presents the possibility that the environment of the organism may have changed during replication altering mutation patterns, potentially as a result of varying levels of stress. Additionally variation around the genome in nucleotide content or mutation patterns could be a result of variation in the dNTP pools available during the replication cycle an excess of any one dNTP can promote its incorporation and inhibit proofreading of the mismatched base (Kunz and Kohalmi 1991), in the case of fast-growing organisms like *E. coli* it is possible that there may be a slight prevalence of dATP within the cell as the other dNTPs are consumed resulting in a mutation bias towards A(&T) closer to the terminus. However it is likely that the forces differentiating the origin and terminus of replication are a combination of selective and mechanistic effects.

The ability to detect and characterise the EcC window shows that this approach has the necessary resolution to identify relatively small regions of the genome (in this case 50 to 90 genes) and provide sufficient detail to begin to tease apart the evolutionary mechanisms and processes at work on these regions as opposed to neighbouring regions or the rest of the genome.

Chapter 6 – Simulated Evolution of Genomic AT Content

6.1 - Introduction

6.1.1 – Factors affecting Genome Nucleotide Composition

There are many processes governing the nucleotide composition of bacterial genomes, broadly classified as; mutation biases, the action of purifying selection and changes in the balance of the two induced by environmental factors associated with the niche occupied.

Mutational processes are inherently biased towards the generation of A or T nucleotides in the place of G or C (Ochman 2003; Lind and Andersson 2008); there are three specific mutations which underpin this bias: deamination of Cytosine resulting in a Uracil, the deamination of 5-methyl-Cytosine to Thymine (Duncan and Miller 1980) and the oxidation of Guanine to 7,8-dihydro-8-oxo-Guanine (Michaels and Miller 1992) which mispairs with Adenine, the latter accounting for approximate 91% of observed mutations during a study of mutational biases in *Salmonella typhimurium* (Lind and Andersson 2008).

In concert with the mutational bias towards AT enrichment there is a selective bias which modulates the AT content of the genomes to maintain or optimise aspects of the of the proteome, such as the hydrophobicity of encoded amino acids (Banerjee, Gupta et al. 2005) and the secondary structure of encoded proteins (Gupta, Majumdar et al. 2000). Additionally there will also be a bias towards the use of preferred codons in order to maintain the efficiency of mRNA translation (Vladimirov, Likhoshvai et al. 2007).

Adoption of a specific niche can also affect nucleotide composition such that GC content is more similar between residents of the same niche than between more closely related species in different niches (Foerstner, von Mering et al. 2005). It has been observed that the adoption of an aerobic lifestyle results in more GC rich genomes (Naya, Romero et al. 2002), likely to counter the increased GC to AT mutation rates expected in an oxygen rich environment, there is also a slight correlation between growth temperature and genome nucleotide composition in some prokaryotes (Wang, Susko et al. 2006).

As previously mentioned (Chapters 1, 3, 4 & 5) the adoption of a pathogenic, symbiotic or intracellular lifestyle is associated with a reduction in effective population size (Mira and Moran 2002) leading to reduced purifying selection and a consequential increase in the AT content of the genome (Moran 2002; Moran, McLaughlin et al. 2009). It is also possible that competition with the host for resources may result in AT enriching mutations being marginally favourable (Rocha and Danchin 2002).

6.1.2 – Static Evolution Simulation Approaches

The principal behind these methods is to 'freeze' evolution and then use the frozen evolutionary trends to predict the equilibrium nucleotide composition of the genomes. In practice this is achieved using a snapshot of evolutionary trends, the polymorphism profile, and altering the base composition of the genome based upon the frequencies of the nucleotide change types observed, in essence projecting the evolutionary 'Heading' of the genomic AT content (see Red line, figure 6.1.3a).

Two different methods for implementing this approach were derived; initially a method based upon stochastic introduction of nucleotide changes one at a time, using the frequencies of the 12 SNP types, was developed. The probability of the application of a given SNP being limited first by the relative abundance of the originating nucleotide (e.g. A in A→C) and then by the frequency of each of the appropriate SNP types (e.g. for A; A→C, A→G & A→T). Whilst this approach applies nucleotide changes in a method similar to that which would be expected *in vivo* i.e. the genome sustains nucleotide changes/mutations individually rather than 'en masse', the polymorphism profile only reflects the observation frequencies of SNPs (potentially including reversions) rather than the frequencies at which they occur and so may lead to errors if applied SNP by SNP.

To that an improvement, based upon a matrix of all possible 16 nucleotide 'changes' including instances of an absolutely conserved base as an N→N 'change', was developed. The current nucleotide composition is treated as a vector and the matrix containing the rates of change (calculated from the SNP & conserved nucleotide counts) is applied as a transformational matrix. In this model there is no assumption as to

whether a base changes or not and there is no reliance upon the generation of random numbers. Additionally, the application of the matrix to the vector is far more computationally efficient than the equivalent number of stochastic nucleotide changes.

6.1.3 – Dynamic Evolution Simulation Approach

The dynamic method is a direct extension of the matrix based static evolution method above, however the matrix of nucleotide changes is allowed to evolve along with the sequence. The evolution of the matrix is based upon the matrices observed from taxon exclusion derived 'time points' for each taxon; for each of the sixteen matrix elements a regression with respect to Log Divergence "Time" was calculated and this was then used as a trajectory to evolve that element forward, based upon the number of observable SNPs generated by the simulation, resulting in a projected evolutionary 'Path' of the genomic AT content (see Blue line, figure 6.1.3a).

The number of SNPs used to calculate the evolved matrix values is based upon the starting number of SNPs and the number of new SNPs applied by the simulation, correcting for the occurrence of multiple hits, which would reduce the number of observable SNPs (see Methods: 2.11.3), resulting in the number of observable SNPs reaching an asymptote around the number of nucleotide sites in the taxon. The observable SNP count is in turn used to calculate the Log Divergence "Time" and derive the corresponding new matrix values.

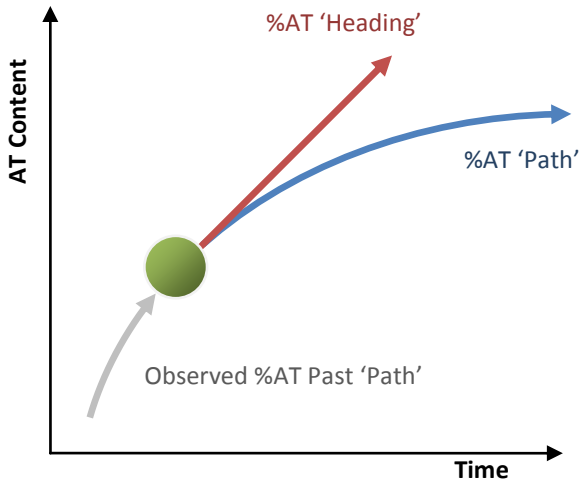


Figure 6.1.3a – An Illustration of the difference between the Static Simulation Approaches - %AT 'Heading' (Red) and the Dynamic Simulation Approach - %AT 'Path' (Blue).

6.1.4 – Aims & Conclusions

In this Chapter I aim to examine three different simulation approaches, two employing a static evolutionary model and one a dynamic evolutionary model, comparing their efficiency and effectiveness. In addition I will use these simulation methods to make a preliminary examination of the projected longer-term consequences of the observed polymorphism profiles on the overall AT content of their respective core genomes.

I observe patterns in the simulated equilibrium AT content of the *Shigellae* and *E. coli* which are consistent with evolutionary differences previously observed (Ch3-5), the differences are more pronounced using the dynamic matrix method for simulation. I observe that complete genome sequences with similar AT content to the equilibrium values show differences in genome size and lifestyle between the *E. coli* and *Shigellae* with organisms showing greater genome degradation and lifestyles which restrict effective population size having AT content closer to the *Shigella* equilibrium values.

6.2 – Static Evolution Simulations

6.2.1 – Stochastic SNP Simulation Approach

As can be seen from figures 6.2.2a & b this simulation approach takes approximately 7,000,000 random nucleotide changes (iterations) in order for the genomic AT content to reach equilibrium. On the computer system used to run the analysis this equated to approximately 48 seconds of CPU time per taxon, whilst this is not an unrealistic compute time larger datasets would require considerable runtime.

There is a notable difference in the trajectories of the three mean values calculated; the *Shigellae* showing the highest equilibrium AT content, then the *E. coli* and the lowest equilibrium values belong to the internal branches. These results are in line with what was observed in Chapters 3 & 4 in terms of the ratio of AT enriching to GC enriching SNPs. The *Shigellae* having the highest value can be explained as a consequence of reduced purifying selection resulting from their lifestyle and the Internal branches having the lowest values is a result of their polymorphism profiles representing 'older' SNPs which have undergone a greater amount of purifying selection than either the *E. coli* or *Shigellae*, purging the more deleterious AT enriching SNPs.

The equilibrium value for each internal branch or taxon is shown below (table 6.2.2a) the equilibrium value is calculated as the mean of the last 10 sample points of the simulation in order to allow for fluctuations associated with the simulation.

Taxon / Branch ID	Divergence Time	Observed AT Content	Equilibrium AT Content
EcA	3.3570	46.9%	54.5%
EcB	3.4408	46.9%	53.5%
EcC	3.4471	47.1%	53.1%
EcD	3.8194	47.0%	52.7%
EcE	3.4048	47.1%	52.7%
Sb	3.3207	47.0%	54.7%
Sd	3.6379	47.0%	53.4%
Sf	3.4645	47.0%	54.8%
Ss	3.2571	47.0%	53.9%
iA	3.6821	46.8%	54.4%
iB	3.8477	46.8%	53.7%
iC	3.9602	46.7%	52.6%
iD	3.9256	46.7%	52.6%
iE	4.2016	46.7%	51.1%

Table 6.2.1a – The Divergence "Time", Equilibrium and Observed AT content of each Taxon / Internal Branch. Internal Branch starting AT content was calculated from the PAML derived sequence/

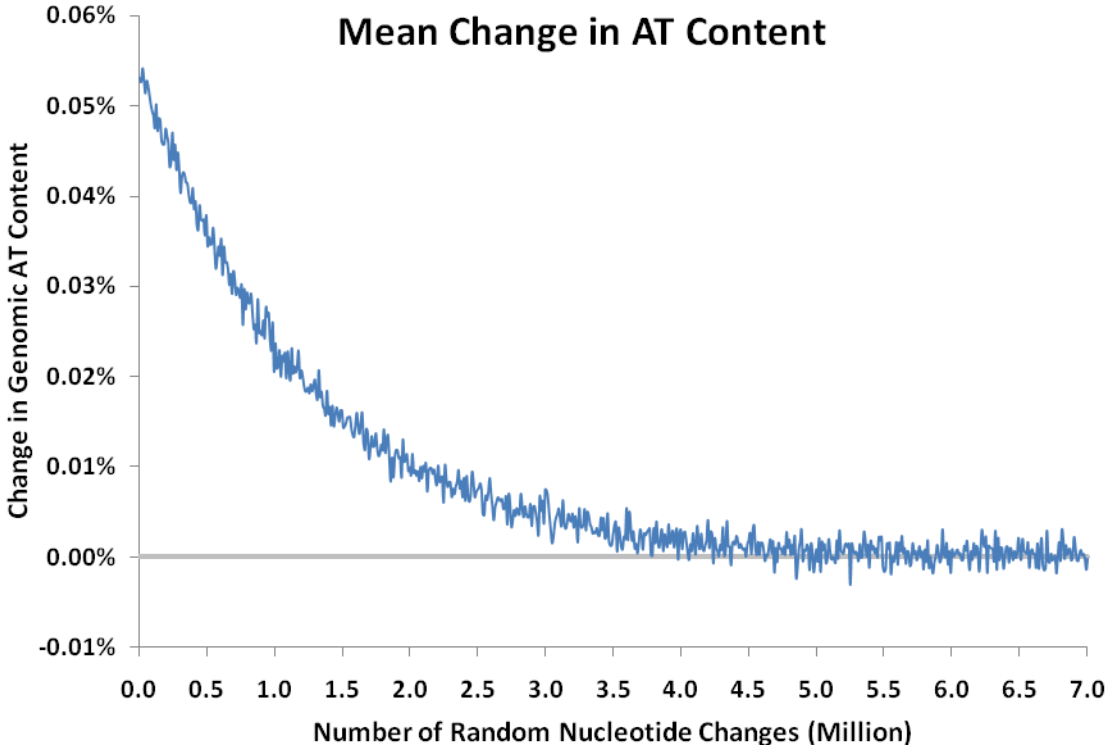


Figure 6.2.1a – Mean trajectory of the difference between successive simulation sample points (10,000 mutations apart).

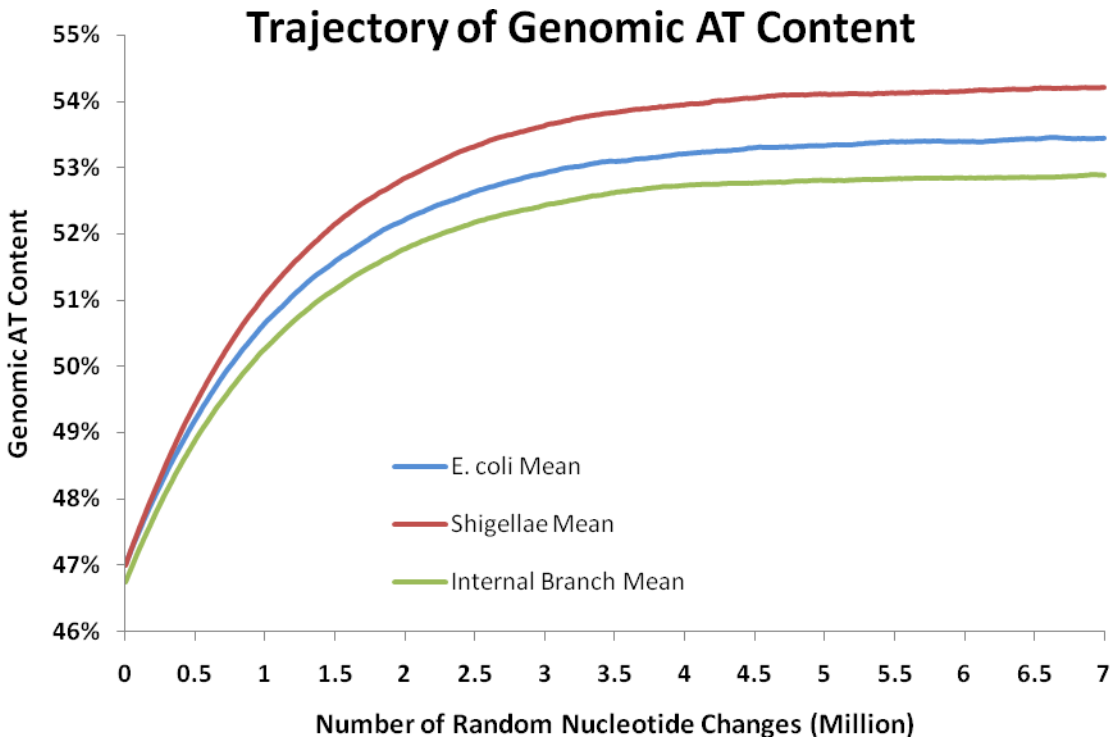


Figure 6.2.1b – Mean trajectory of genomic AT content for each of the *Shigellae*, *E. coli*, and Internal Branches as the simulation progresses.

As can be seen below there is a significant relationship between the divergence time associated with a taxon or internal branch and the equilibrium AT content of the corresponding sequence, representing a time dependant trend away from an AT rich genome.

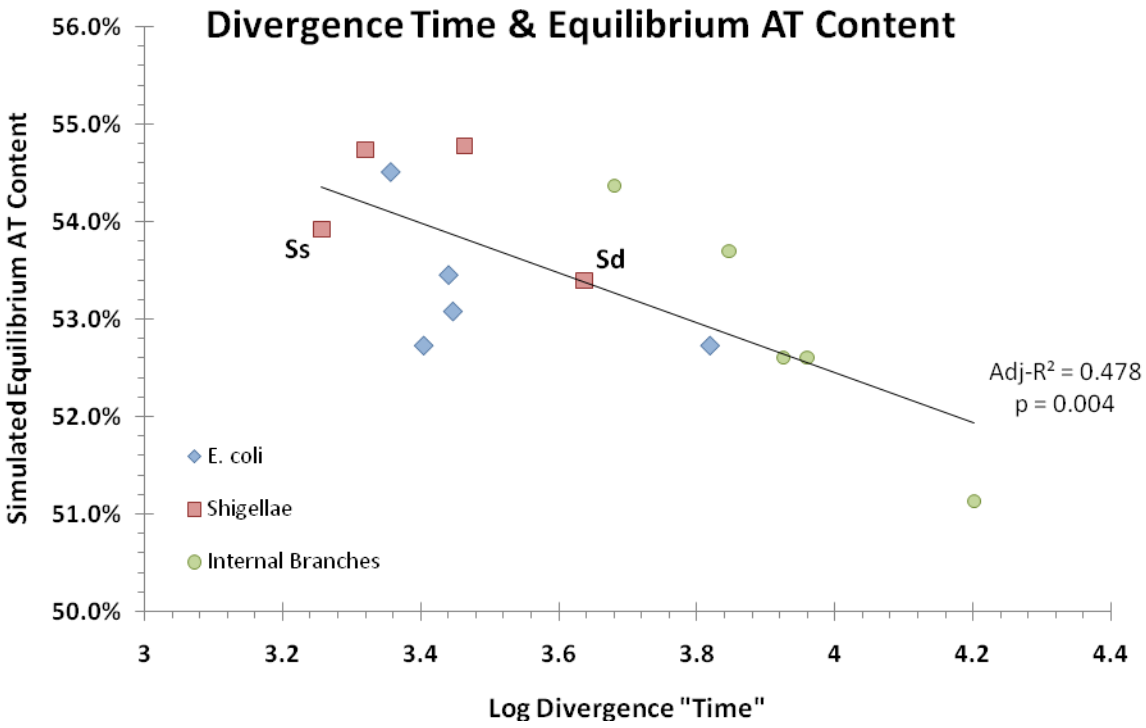


Figure 6.2.1c – The simulated equilibrium AT content of the taxa and internal branches versus Log divergence time, showing the time dependence of equilibrium AT content.

Whilst there is variation in the equilibrium AT content based upon different 'timepoints' of each taxon, derived using taxon exclusion, the differences between the *Shigellae* and *E. coli* are still evident. Interestingly both *S. sonnei* and *S. dysenteria*e show more *E. coli* – like equilibrium values than the other *Shigellae* (figures 6.2.1c & d), this is in contrast to the previously established pattern (Chapters 3 & 4) of *S. sonnei* being the outlier of the *Shigellae* and showing more *E. coli*-like patterns of evolution.

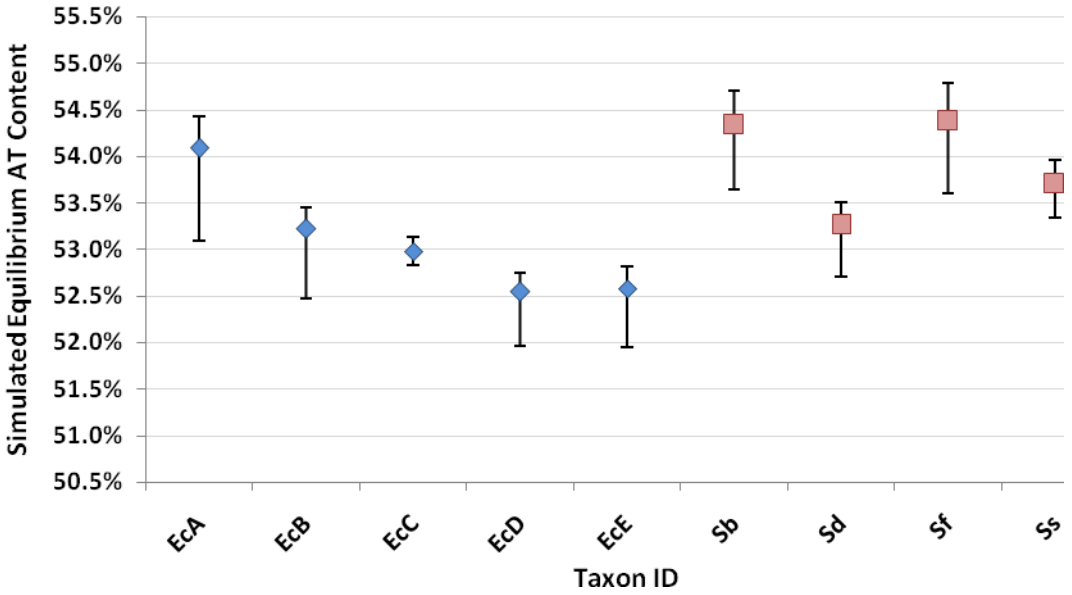


Figure 6.2.1d – The mean value and spread of values for simulated equilibrium AT contents for each taxon based upon all the polymorphism profiles obtainable through taxon exclusion.

6.2.2 – Static SNP Matrix Simulation Approach

The matrix based approach reaches the equilibrium AT content value within 4000 iterations (Figures 6.2.2a & 6.2.2b), which is substantially lower than the 7,000,000 iterations necessary in the stochastic SNP approach. In terms of runtime necessary running the matrix approach for 100,000 iterations (sampling every 100) required approximately 1.4 seconds of CPU time per taxon, however the simulation reached equilibrium at around iteration 3000 yielding a runtime to equilibrium of approximately 0.04s per taxon, this gives an approximately 1200-fold reduction in compute time. Likely a consequence of the matrix approach implementing the equivalent of 10,000 to 50,000 SNPs per iteration.

As before there is a clear separation in the trajectories of the three mean values calculated; the *Shigellae* showing the largest equilibrium AT content, then the *E. coli* and the lowest equilibrium values belong to the internal branches. The equilibrium values obtained from this simulation method are far higher than those obtained using the stochastic SNP method. This is more likely a limitation of the power of the stochastic simulation than it is an error or bias in the matrix simulation, which makes fewer assumptions.

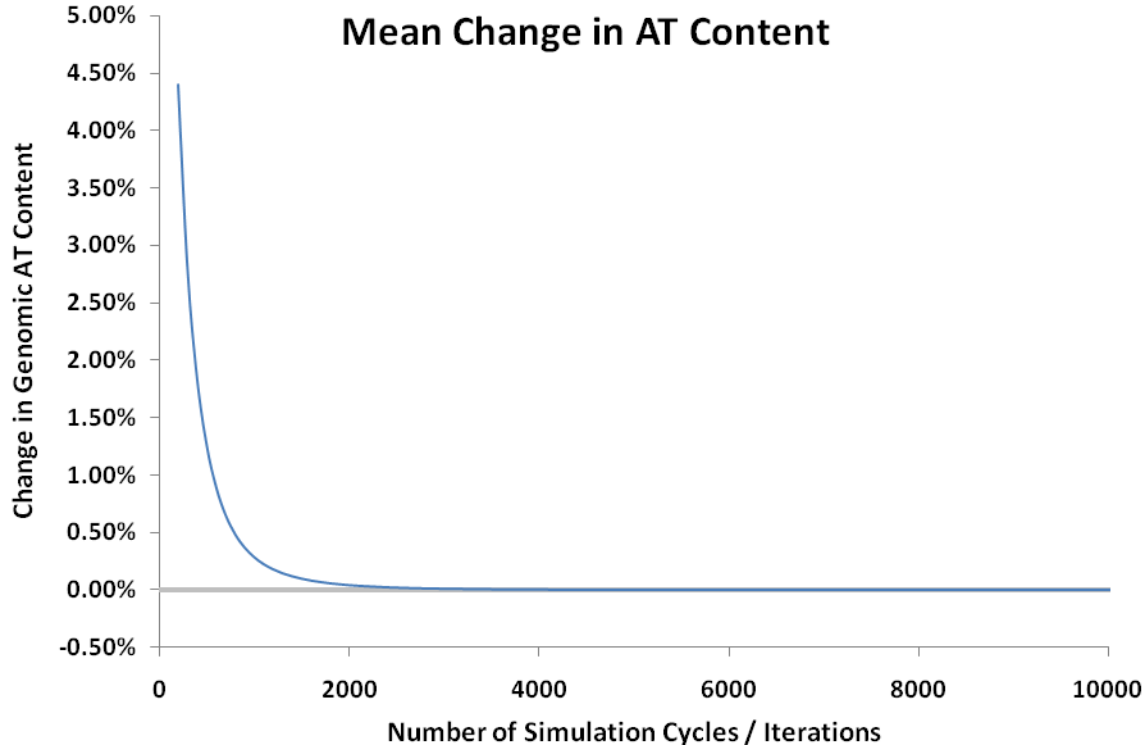


Figure 6.2.2a – Mean trajectory of the difference between successive simulation sample points (100 iterations apart).

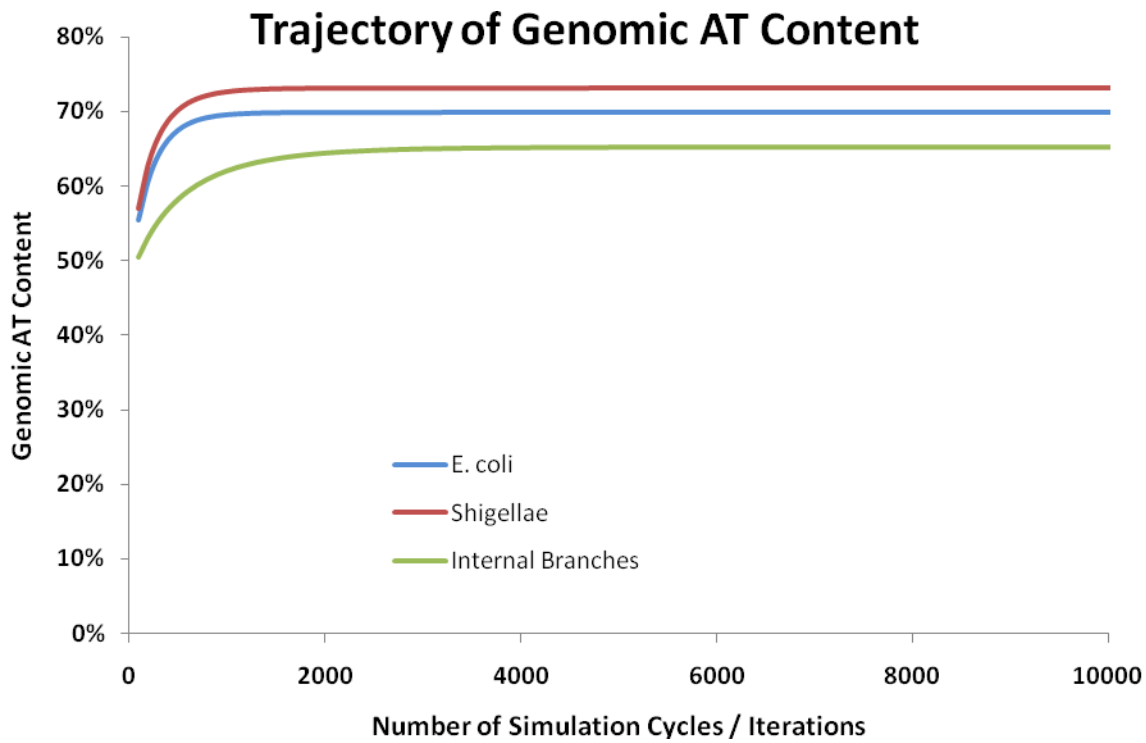


Figure 6.2.2b – Mean trajectory of genomic AT content for each of the *Shigellae*, *E. coli*, and Internal Branches as the simulation progresses.

The equilibrium value for each internal branch or taxon is shown below (table 6.2.2a) the equilibrium value is calculated as the mean of the last 10 sample points of the simulation in order to allow for fluctuations associated with the simulation.

Taxon / Branch ID	Divergence Time	Starting AT Content	Equilibrium AT Content
EcA	3.3570	46.9%	69.9%
EcB	3.4408	46.9%	70.3%
EcC	3.4471	47.1%	68.5%
EcD	3.8194	47.0%	63.3%
EcE	3.4048	47.1%	70.6%
Sb	3.3207	47.0%	74.2%
Sd	3.6379	47.0%	71.3%
Sf	3.4645	47.0%	74.2%
Ss	3.2571	47.0%	72.9%
iA	3.6821	46.8%	70.3%
iB	3.8477	46.8%	67.7%
iC	3.9602	46.7%	66.8%
iD	3.9256	46.7%	65.4%
iE	4.2016	46.7%	55.6%

Table 6.2.2a – The Divergence "Time", Equilibrium and Starting AT content of each Taxon / Internal Branch. Internal Branch starting AT content was calculated from the PAML derived sequence/

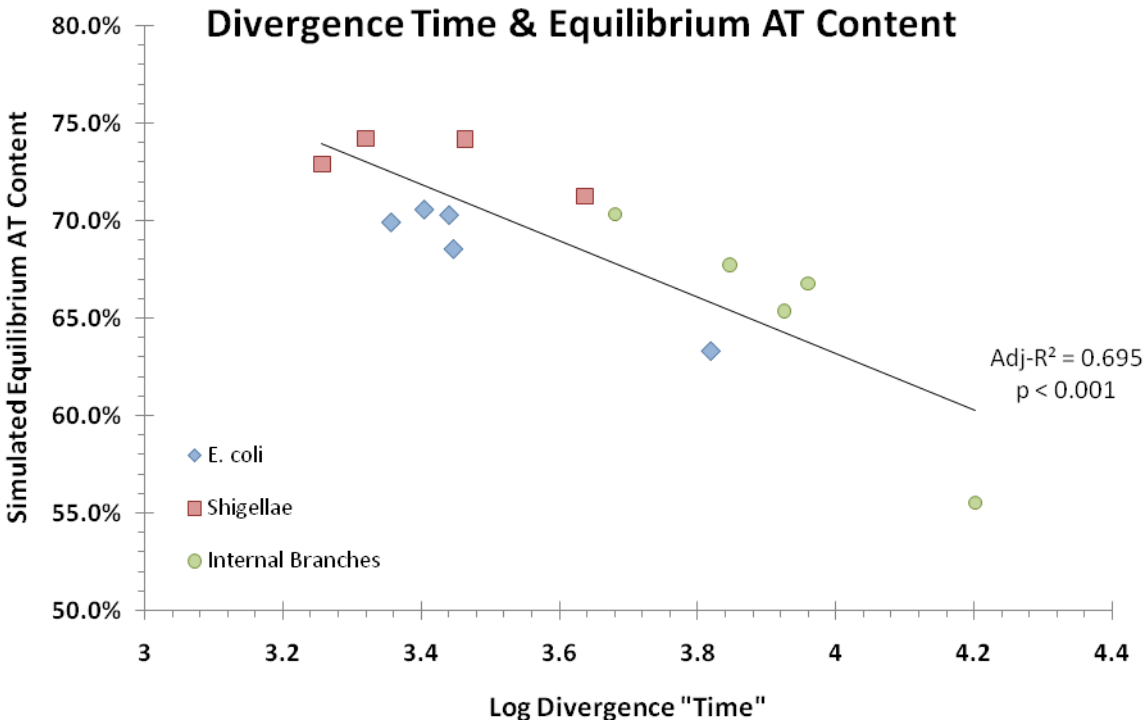


Figure 6.2.2c – The simulated equilibrium AT content of the taxa and internal branches versus Log divergence time, showing the time dependence of equilibrium AT content.

As with the earlier simulation there is a clear time dependence of the simulated equilibrium AT content (figure 6.2.2c), with the relationship being stronger for these simulation results and highly significant. This result also, as expected, agrees strongly with the pattern observed in the +AT/+GC ratio (Chapter 3), reflecting the time dependant purging of the more deleterious AT enriching SNPs, consequently ‘older’ polymorphism patterns represent a trend toward a less AT rich equilibrium value.

6.3 – Dynamic Evolution Simulation

6.3.1 – Multiple Hit Correction Testing

In order to test the probabilistic estimate of the number of SNPs observable when *s* SNPs are applied (at random) to a genome of size *B* nucleotides, test distributions were generated of 1000 replicates of various numbers of SNPs (ranging from 10 to 10,000) were applied to 'genomes' of 1,000 10,000 and 100,000 nucleotides. Figure 6.3.2a, below illustrates the *s*=10,000 & *B*=100,000 distribution.

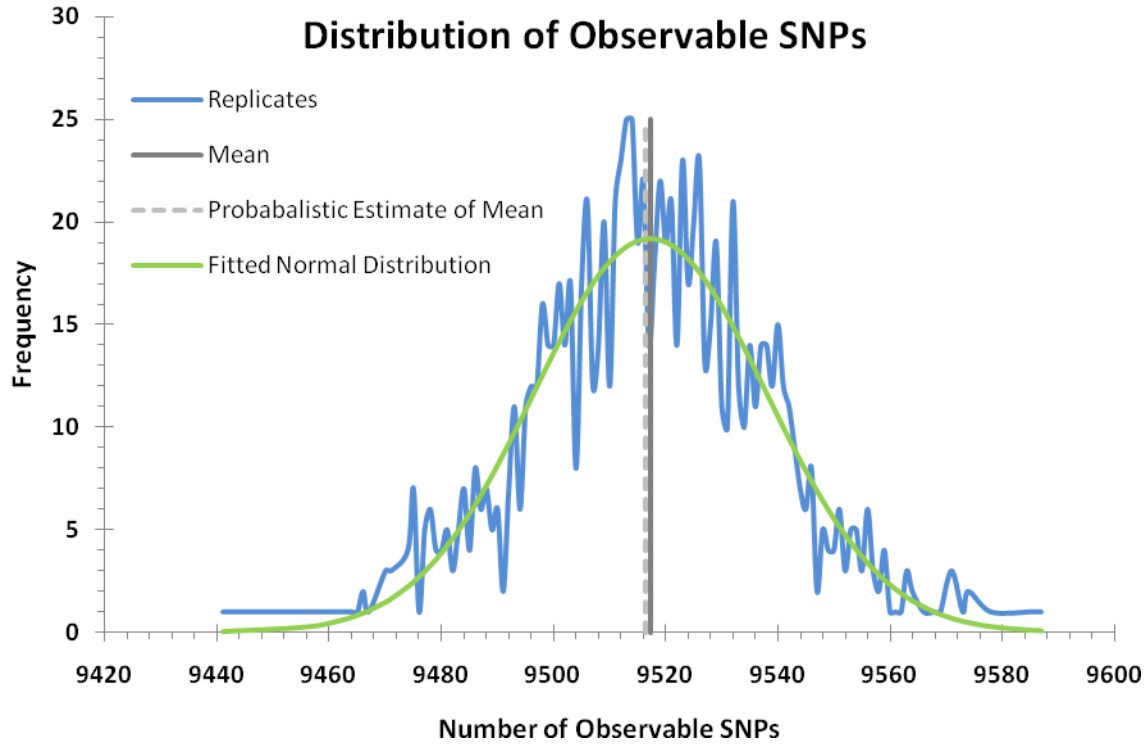


Figure 6.3.1a – Distribution of 1000 replicates of 10,000 SNPs applied to a genome of 100,000 nucleotides, showing the Mean (solid vertical line), Probabilistic Estimate of the Mean (dashed vertical line) and the fitted normal distribution.

Distribution Parameters		Distribution Mean	Estimated Mean	Distribution Log Divergence Time	Estimated Log Divergence Time
B	*s*				
1000	10	9.94	9.96	0.696	0.697
1000	100	95.36	95.21	1.678	1.678
10000	10	10.00	10.00	0.699	0.699
10000	100	99.50	99.51	1.697	1.697
10000	1000	951.44	951.67	2.677	2.677
100000	10	10.00	10.00	0.699	0.699
100000	100	99.94	99.95	1.699	1.699
100000	1000	992.97	995.02	2.696	2.697
100000	10000	9517.23	9516.30	3.677	3.677

Table 6.3.1a – The Distribution and Estimated mean number of Observable SNPs and corresponding Log Divergence Time for the distributions of observable SNPs for several combinations of Genome size (*B*) and number of applied SNPs (*s*).

Whilst there is a small amount of difference between the distribution mean number of observable SNPs and the probabilistic estimate of the number of observable SNPs, the differences are insignificant (paired t-test of the Distribution Mean and Estimated Mean; p = 0.616). This is also clear when considering the Log Divergence times from the two methods which differ by a maximum of 0.001 (table 6.3.1a).

It is also important to consider the computational cost of the two approaches; the probabilistic estimate requires less than a second of compute time, whereas calculating the mean of the distribution of 1000 applications of *s* SNPs to *B* nucleotides takes 73.4s for 10,000 SNPs onto 100,000 nucleotides. Application of the distribution approach to a realistic example, in this case EcA, requires the application of 4550 SNPs onto 2,063,004 nucleotides which for 1000 replicates would take approximately 25 minutes of compute time. Given that this method has to be implemented once per iteration of the simulation and therefore thousands of times per taxon the distribution approach is impractical.

6.3.2 – Dynamic Simulation

The number of observable SNPs becomes saturated at 100% of the number of nucleotides in the taxon between 75 and 200 iterations into the simulation, depending on the taxon (figure 6.3.2a). Consequently the matrix becomes static after this point as the regression used to extrapolate the matrix values is based upon the Log Divergence Time, which is in turn based upon the number of observable SNPs.

Additionally the rate of change reveals that the equilibrium AT content value for the vast majority of taxa is reached after approximately 60 – 80 iterations, all taxa having equilibrated by 100 iterations (figure 6.3.2b). Given this equilibration point and the runtime for 1000 iterations of 0.99s this gives an effective runtime to equilibrium of 0.011s per taxon, approximately four times as fast as the static matrix approach and nearly 4500 times as fast as the stochastic method.

As seen in previous simulations the *Shigellae* show a higher average equilibrium AT content than do *E. coli*, reflecting the lifestyle differences and associated difference in selection between the *E. coli* and *Shigellae.*

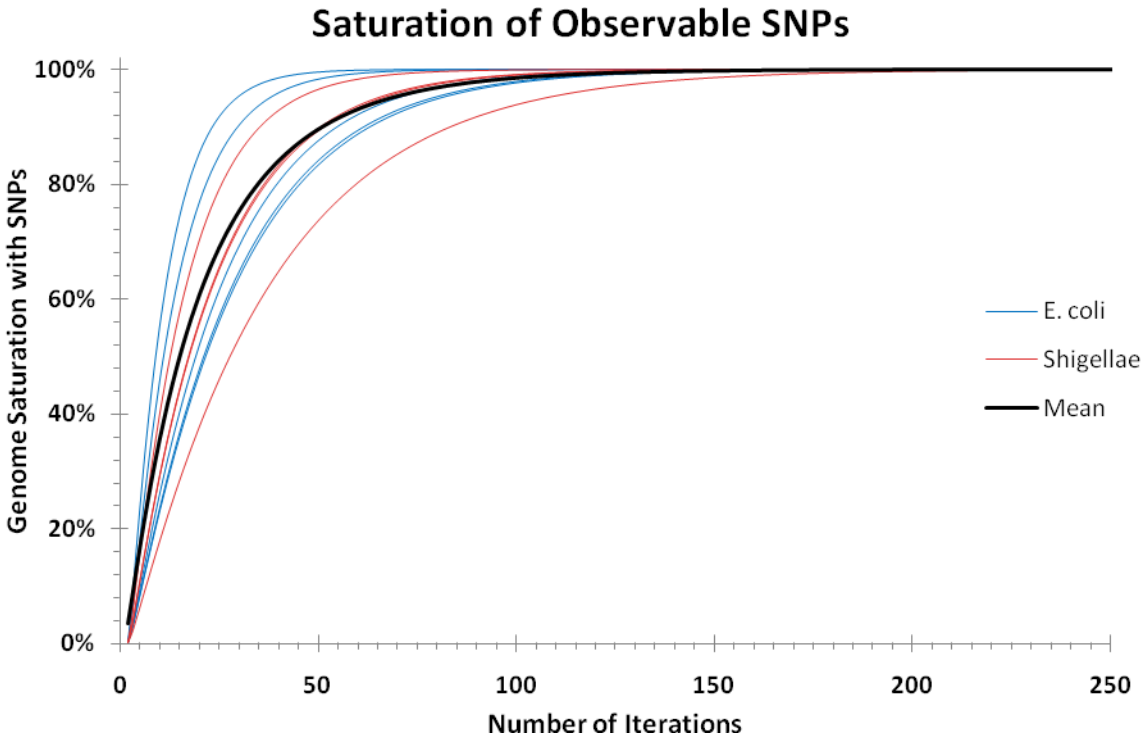


Figure 6.3.2a – Saturation of the Genome with Observable SNPs against simulation iteration, showing both the mean and the individual traces for each of the 5 *E. coli* and 4 *Shigellae*. Only the first 250 iterations are shown.

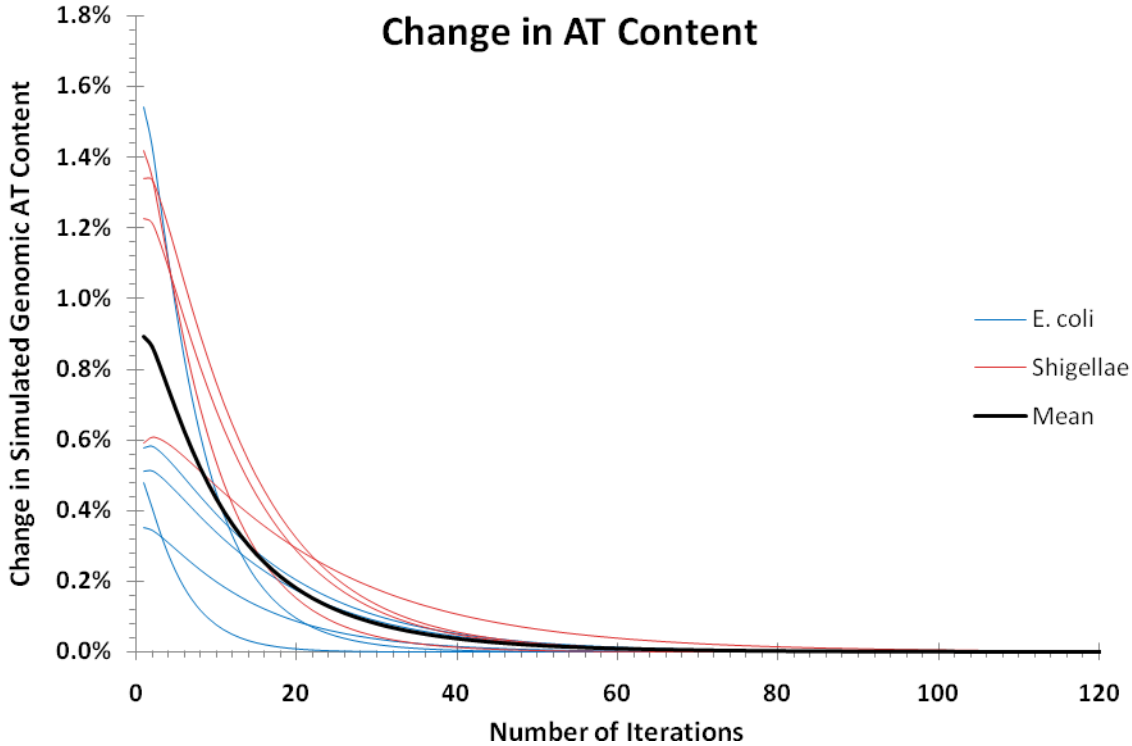


Figure 6.3.2b – Mean and per taxon differences between the AT content of successive simulation sample points, which are 1 iteration apart. Only the first 120 iterations are shown

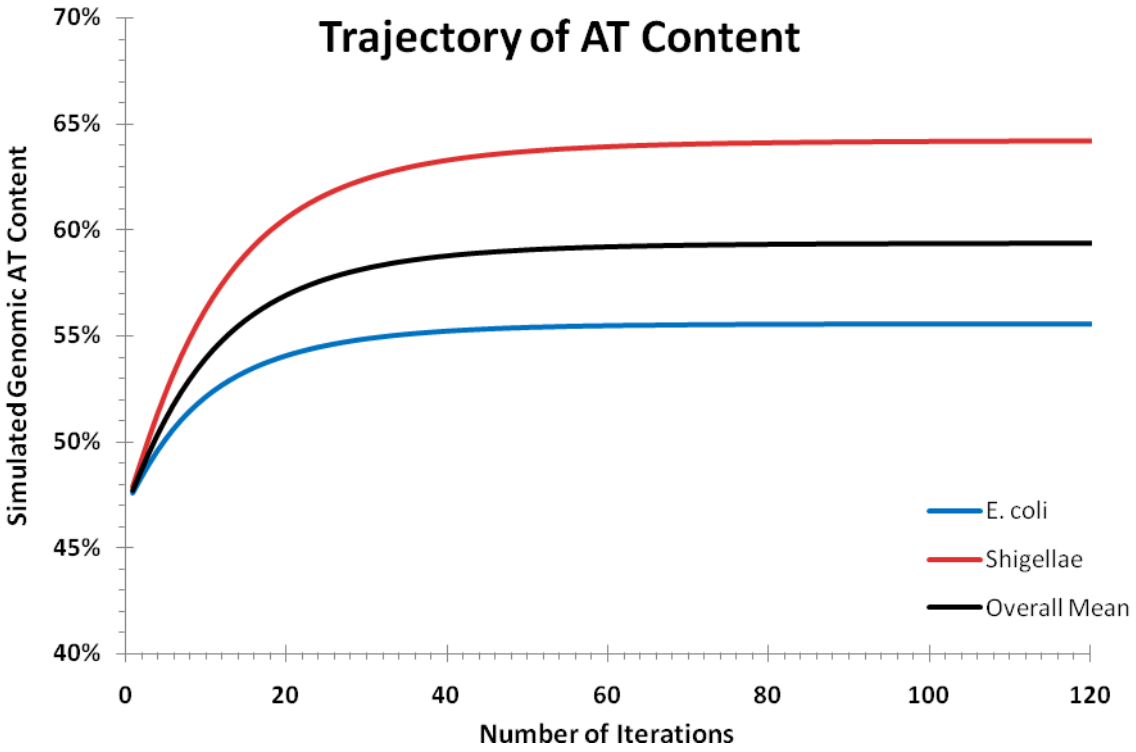


Figure 6.3.2c – The trajectory of the Mean AT content of the *Shigellae* and *E. coli* as well as the Overall Mean. Only the first 120 iterations are shown.

The equilibrium value for each internal branch or taxon is shown below (table 6.3.2a) the equilibrium value is taken as the value after 1000 iterations as the simulations all showed no change (to 6 decimal places) for the last 250 iterations.

Taxon / Branch ID	Divergence Time	Starting AT Content	Equilibrium AT Content
EcA	3.3570	46.9%	58.0%
EcB	3.4408	46.9%	52.3%
EcC	3.4471	47.1%	56.7%
EcD	3.8194	47.0%	50.1%
EcE	3.4048	47.1%	60.5%
Sb	3.3207	47.0%	67.3%
Sd	3.6379	47.0%	61.8%
Sf	3.4645	47.0%	65.3%
Ss	3.2571	47.0%	62.3%

Table 6.3.2a – The Divergence "Time", Equilibrium and Starting AT content of each Taxon

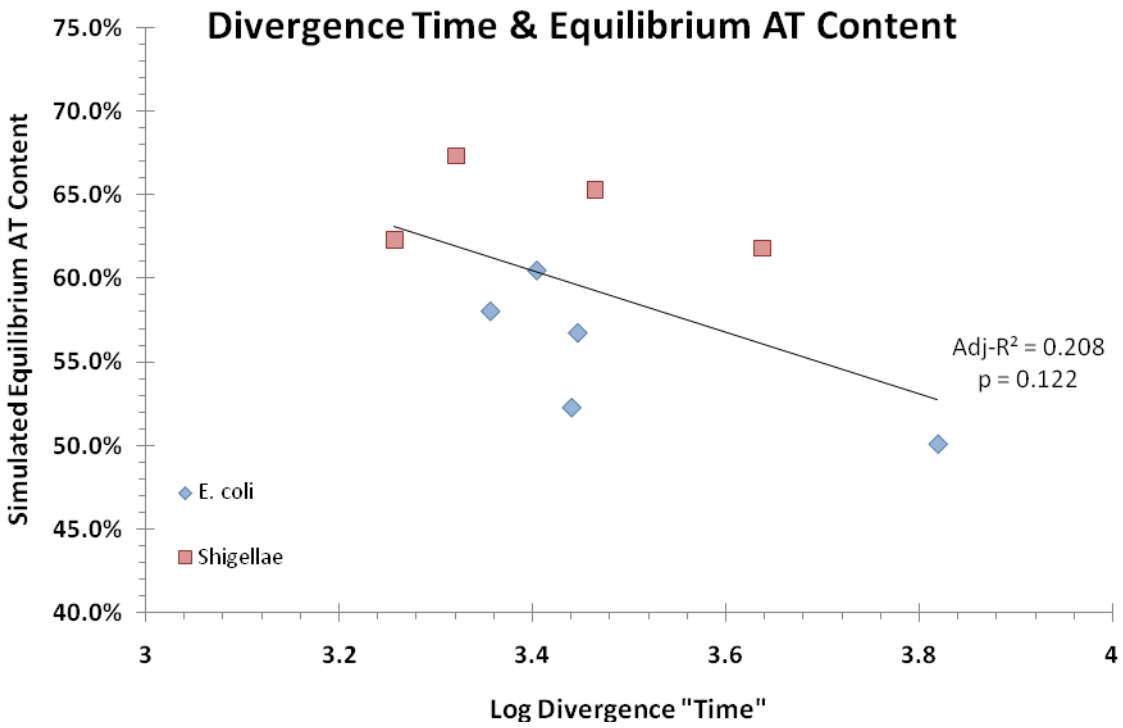


Figure 6.3.2d – The simulated equilibrium AT content of the taxa versus Log divergence time

Whilst there appears to be a trend in these values with log divergence time (figure 6.3.2d), the regression is not statistically supported, this is partially the result of the exclusion of the internal branches from the analysis, as there is no acceptable method of inferring 'timepoints' for several of the internal branches. The *Shigellae* show higher equilibrium AT content values than all of the *E. coli*, again likely reflecting their intracellular lifestyle and the resulting reduced purifying selection.

6.4 – Summary of Results

- The stochastic static evolution simulation reached equilibrium at approximately 7 million iterations with there being a clear difference between the *E. coli* and *Shigellae* mean equilibrium AT content.

- The equilibrium values range from 51 to 55% AT content, and show a clear and significant downward trend with divergence time.

- The matrix static evolution simulation reached equilibrium at approximately 3,000 iterations again showing a clear difference between the mean equilibrium AT content of the *Shigellae* and *E. coli*.

- The equilibrium values from this simulation method are between, 55 and 74% AT content (63 & 74%, excluding internal branches) and also show a clear and significant downward trend with divergence time.

- Multiple hit correction using the probabilistic estimate provides an accurate and reliable approximation of the mean of the distribution of number of observable SNPs.

- The dynamic matrix evolution simulation reached equilibrium for all taxa, between 100 and 120 iterations, with the mean equilibrium AT content showing clear separation of the *E. coli* and *Shigellae*.

- The equilibrium values for this simulation range from 50 to 67% AT content, internal branches were not included due to difficulties in deriving multiple SNP matrices for them. There appears to be a divergence time dependent trend for the equilibrium AT content values, however it is not significant.

6.5 - Discussion

6.5.1 – Comparison of Methods

The static methods both show a clear and significant divergence time dependent trend in the equilibrium AT content of the genomes, which separate the *Shigellae* from *E. coli*. The similarity between the two approaches largely ends there, with the stochastic approach showing far lower equilibrium values than the matrix-based approach.

These lower values also show a much smaller separation than do the matrix method values, this is likely a consequence of the assumption that all nucleotide selections will result in a base change and the consequent exclusion of information from the sequence data, specifically the exclusion of the number of occurrences of an absolutely conserved base, which in the simulation would correspond to an unchanging base. This information is vital when attempting to determine the rate of change for any given base, as whilst two nucleotides may have similar distributions of their 'changes' one may be significantly more likely to be conserved than the other and thus should display lower rates of change.

Additionally compute time for the two methods is considerably different, with the stochastic method taking on average 34 times longer (in terms of CPU time) to reach equilibrium than the matrix method. This is partially due to the matrix method implementing the equivalent of 10,000 to 50,000 SNPs per iteration and as a consequence even though it reaches equilibrium in only ~3000 iterations this represents the application of approximately ten times the number of SNPs as the stochastic method.

A comparison of the two approaches is somewhat more complicated as they reflect slightly different questions, the static models simulating the evolutionary 'direction' inferred from the current observed SNP profile and the dynamic model simulating the evolutionary 'path' inferred from the observed rates of change of the SNP profiles for each taxon.

However at least in terms of computational efficiency the dynamic matrix algorithm is faster still than the static matrix algorithm, suggesting that the increased complexity of each iteration is more than counterbalanced by the decreased number of iterations necessary to reach equilibrium value, resulting in the algorithm being four times faster (in compute time) to reach equilibrium.

Method	Advantages of Method	Disadvantages of Method	Mean CPU Time to Equilibrium Per Taxon
Stochastic	Mimics the biology, i.e. one SNP at a time	Does not allow for multiple hits or unchanged bases this excludes differences in rate of change between the bases.	48s
Static Matrix	Includes information on unchanged bases and therefore differing rates of change between the nucleotides	Assumes evolution is a static process.	0.04s
Dynamic Matrix	Allows patterns of evolution to change over time.	Relies on inferring past evolutionary patterns and assumes that evolution will progress along that trajectory	0.011s

Table 6.5.1a – Summary of the benefits and drawbacks of each of the simulation approaches and the mean CPU runtime, per taxon, to reach equilibrium.

Based upon table 6.5.1a (above) and the earlier discussion it is fairly clear that the simulation strategies based upon the use of matrices of nucleotide interchange frequencies are the most efficient and suffer from the fewest drawbacks.

6.5.2 – Static Evolution (Matrix method) Results

The values of the simulated equilibrium AT content for the *Shigellae* and *E. coli* show slightly different ranges, the *E. coli* being between 63% & 71% and the *Shigellae* between 71% and 75%. The higher range in the *Shigellae* likely reflects the higher +AT/+GC ratio consistently observed (Chapters 3, 4 & 5) and therefore is result of their reduced effective population size and consequentially reduced purifying selection.

It is interesting to compare the resulting equilibrium values of the simulations to the known genomic compositions of other proteobacteria, specifically the list of the 428 completed genome projects represented in the Centre for Biological Sequence Analysis's (CBS)

Genome Atlas Database (Hallin and Ussery 2004). For the *E. coli* range this corresponds to various species, from the cytoplasmic parasite of insects *Wolbachia* at ~64% AT (Wu, Sun et al. 2004) through the human pathogen *Francisella* at ~67.7% AT (Francis 1921) to the obligate intracellular parasite *Rickettsia* with an AT content of ~71% (Andersson, Zomorodipour et al. 1998). The *Shigellae* equilibrium values correspond to a range of genomes which also includes *Rickettsia* as well as the obligate intracellular pathogen of humans (and dogs) *Erlichia* (Dumler, Madigan et al. 2007) at ~72.5% AT and the endosymbiont of aphids, *Buchnera aphidicola* with an AT content of 73% to 75% (Perez-Brocal, Gil et al. 2006). Overall both groups show AT contents heading towards that observed in a mix of pathogenic and parasitic bacteria, there being a slightly greater bias towards intracellular lifestyles in the group corresponding to the *Shigellae* AT content heading.

The mean size of the proteobacteria with similar nucleotide composition to the equilibrium values for the *Shigellae* and *E. coli* are different with the genomes similar to the *Shigellae* being approximately 0.5Mbp smaller (mean sizes of 1.57Mbp and 1.08Mbp for genomes corresponding to the *E. coli* and *Shigellae* equilibrium AT content respectively). This reflects a trend in the *Shigellae* of greater genome degradation as compared to the *E. coli* and hints at a future genome which is smaller and more AT rich than that of the *E. coli*.

The correlation ($r = 0.965$, $p < 0.001$) of the equilibrium values with the divergence time reflects the previously observed time dependence of the +AT/+GC ratio (Chapters 3, 4 & 5), and is expected given that it is this bias which, under the static model, determines the 'direction' or 'heading' of the genomic AT content.

6.5.3 – Dynamic Evolution Results

The simulated equilibrium AT content for the *Shigellae* and *E. coli* again show different ranges, with the *Shigellae* between 62% & 67% and the *E. coli* between 50% & 60%, the latter representing minimal change from the current AT content of ~49%. As with the static simulation the higher AT content of the *Shigellae* is likely a consequence of the previously observed (Chapters 3, 4 & 5) effects of reduced purifying selection.

A comparison of these ranges to the published proteobacteria genomes reveals a slightly stronger differentiation in the bacteria corresponding to the *E. coli* and *Shigellae* equilibrium values. The former showing similarity to the human pathogens; *Yersinia pestis* (Oliver 2005) with an AT content of ~53% and *Vibrio vulnificus* (Perry and Fetherston 1997) with an AT content of ~58% and the respiratory pathogen of pigs *Actinobacillus pleuropneumoniae* (Dom, Haesebrouck et al. 1994) with an AT content of 62%. The *Shigellae* show similarity to both the human pathogen *Haemophilus influenzae* (Hogg, Hu et al. 2007), to intracellular parasites (of insects) such as Wolbachia (Wu, Sun et al. 2004) and to an intracellular enteropathogen of pigs *Lawsonia intracellularis* (McOrist, Gebhart et al. 1995). Overall there is a trend for the equilibrium AT content in the *E. coli* to reflect that of pathogenic bacteria with extracellular modes of infection as with the examples above, whilst the *Shigellae* show equilibrium AT contents more in line with bacteria with intracellular lifestyles as well as some pathogenic bacteria.

In addition to the functional differentiation of the two groups of AT content matches, there is a marked difference in genome size of the two groups; the genomes corresponding to *E. coli* equilibrium AT content have a mean genome size of 3.90Mbp whereas those corresponding to *Shigellae* equilibrium AT content have a mean genome size of 1.85Mbp, representing the previously documented reduction of genome sizes in intracellular bacteria (Andersson and Kurland 1998; Sallstrom and Andersson 2005).

The correlation of the equilibrium values for the taxa and the +AT/+GC ratios from the 'extant' polymorphism profiles (as opposed to those inferred through the use of taxon exclusion) is both strong and significant ($r = 0.910$, $p = 0.001$). This indicates that there is still a strong relationship between the inferred equilibrium AT content and the +AT/+GC ratio observed, even when the polymorphism profile is allowed to evolve with time along with the genome nucleotide composition.

6.5.4 – Differences in simulated AT content 'Headings' and 'Paths'

In essence the static matrix simulation infers the 'heading' of the AT content evolution of a particular taxon, whilst the dynamic simulation infers the more complicated evolutionary 'path' the taxon's AT content will take, allowing for changes in direction as the polymorphism profile is moderated by selective pressures. Interestingly both the *Shigellae* and *E. coli* show similar evolutionary 'Headings' with a small difference in the mean projected AT content from the static simulation (3.4%) and a similarly small difference in the mean size of the proteobacteria with similar AT contents (0.49Mbp), by contrast the differences observed in the 'Paths' i.e. the mean equilibrium AT content values from the dynamic simulation are larger (8.7%) and the difference in the genome size of the proteobacteria with corresponding AT content is far greater (2.05Mbp).

This increased difference in the dynamic simulation is mainly the result of a large shift in the *E. coli* results, whilst the *Shigellae* results also shift they do so by a far smaller amount (an average of -9% AT for the *Shigellae* and -14.3% AT for the *E. coli*). The shift in the equilibrium values between the two simulations reflects the effects of selection (in this case purifying – see Chapter 3) on the projected equilibrium AT content of the taxa, as selection is a time dependant process. Consequently the reduced effective purifying selection in the *Shigellae* manifests as a smaller difference between its 'Heading' and its 'Path' (see differences illustrated earlier in figure 6.1.3a) i.e. there is little moderation of the AT enrichment bias with evolutionary time.

Chapter 7 – Overall Discussion

7.1 – Aims & Results

The aim of this project was to examine and to characterise the patterns and trends of evolution over short timescales or between closely related organisms and to examine if and/or how these trends vary with population structures or lifestyles. These objectives were based upon prior work by Rocha et al (2006), whereby values of d*N*/d*S* were observed to be dependent on divergence time.

7.1.1 – Time Dependence

In line with the observations by Rocha et al (2006) d*N*/d*S* values were found to be divergence time dependent, with a trend over time towards lower values within the *E. coli* – *Shigellae* dataset (fig. 3.4.1a), reflecting gradual purging of nonsynonymous changes. This time dependence was also observed in the distribution of SNPs; showing a clear enrichment for SNPs at more degenerate site types with increasing divergence time i.e. a preference for fourfold degenerate sites over non-fourfold degenerate sites (fig. 3.3.3a) and a preference for the codon positions as follows; $3^{rd} >> 1^{st} > 2^{nd}$ (figs 4.2.2a & b).

This time dependence of SNP purging applies not only the location of the SNP but also to the type of SNP as well as evidenced by the metric ratios of +AT/+GC and Ti/Tv. The pattern of this time dependence is different for each of the metric ratios; +AT/+GC shows a consistent time-dependent enrichment of +GC SNPs regardless of site type (figs. 3.6.3a-c & 4.3.2a-d) which is also observed in a time dependent trend towards lower simulated equilibrium genomic AT content (figs. 6.2.1c, 6.2.2c & 6.3.2d). In contrast Ti/Tv displays an enrichment of Transitions (Ti) with increased divergence time at non-fourfold degenerate sites (fig. 3.6.4b) and at the 1^{st} codon position (fig. 4.3.3a), whilst showing no time-dependant trend at fourfold degenerate sites (fig. 3.6.4c) and the 2^{nd} and 3^{rd} codon positions (figs. 4.3.3b&c).

The differences in the two metric ratios strongly suggest that they represent two different patterns of selection, with the Ti/Tv ratio capturing the pattern associated with amino acid

sequence, as evidenced by the absence of a time dependant trend at the more degenerate site types (Q-sites and 3rd codon position) as well as at the 2nd codon position where there is no degeneracy. The +AT/+GC ratio shows no variation with codon position or site degeneracy and so likely reflects the pattern of selection acting upon the nucleotide content and codon bias of the sequences independent of the encoded amino acids.

There is also a clear time-dependant trend in the mean metabolic cost per amino acid change, with larger divergence times being associated with a lower meat metabolic cost (fig. 4.4.2a). This reflects the time-delayed action to purge gains of the more costly amino acids, a result of the proteome being poor in 'expensive' amino acids; consequently short term patterns of mutational change are likely to result in a bias towards their gain, this pattern of gain and loss bias is observed in all the taxa in the dataset.

7.1.2 – Lifestyle & Niche Effects

The different lifestyles of the *E. coli* and *Shigellae* included in the dataset provide an opportunity to examine the effects of these lifestyles on evolutionary pressures and processes over the recent evolutionary past. The facultative intracellular lifestyle of the *Shigellae* provides not only a completely different replication environment to the *E. coli* but additionally limits their effective population size due to physical constraints of host cell volume and low number of invading bacteria.

In addition to higher d*N*/d*S* values (fig. 3.4.1a) the *Shigellae* show a markedly reduced bias in the distribution of their SNPs compared to the *E. coli*, the *Shigellae* showing a larger proportion of SNPs at NQ sites (fig. 3.3.3a) and the 1st and 2nd codon positions (fig. 4.2.2a). In all cases the gradient of the regression line (representing the rate of selective purging of the nonsynonymous changes) for the *Shigellae* is lower than that for the *E. coli*, this can also be visualised as a tendency for the two groups to lie on opposite sides of the 'overall' regression line.

With the metric ratios (+AT/+GC and Ti/Tv) there is a difference between the trend, or distance from the overall regression, in the *Shigellae* and *E. coli* wherever there is a time-dependent overall trend in that metric (figs. 3.6.3a-c, 3.6.4a-c, 4.3.2a-d & 4.3.3a-d), so at

sites where there is no observable change in the bias of the ratio with time there is no clear distinction of the two groups of taxa, supporting the conclusion that the difference between the two is due to differing levels of purifying selection rather than mutational differences in the intracellular environment (fig. 3.6.5a,b). The one exception to this is the metabolic cost of amino acid changes, whilst there is a clear divergence time dependent trend towards lower metabolic costs there is no clear distinction between the *E. coli* and *Shigellae* (fig. 4.4.2a). This is likely a consequence of the slower evolution of amino acid changes in concert with the short evolutionary timescales observed and the inherently smaller number of changes observable at the amino acid level.

The consistency of these observed differences varies with position along the aligned core genome; the differences proportion of SNPs observed at any particular nucleotide site type is the most consistent (figs. 5.4.2a,b & 5.4.3a,b). The Ti/Tv ratio differences are most consistent at NQ sites followed by all sites (fig. 5.5.1a-c), whilst the +AT/+GC ratio shows a clear trend of decreasing separation of the *Shigellae* and *E. coli* with distance from the origin (fig. 5.5.3a-c), reflecting the associated reduction in selective constraint.

The differences observed between the simulated equilibrium AT content from the static evolution model and the dynamic evolution model reveal potential markers of the relative difference in the selective pressures acting upon the *Shigellae* and *E. coli*, with the *Shigellae* showing a smaller difference between the two simulation methods than the *E. coli* (tabs. 6.2.2a cf. 6.3.2a), reflecting greater purifying selection in the *E. coli*. Additionally the equilibrium AT content of the genomes is similar to that of other proteobacteria with comparable lifestyles, extracellular pathogen in the case of the *E. coli* and intracellular pathogen or endosymbiont in the case of the *Shigellae*.

7.1.3 – Distance from Origin Effects

Preliminary results also show indications of evolutionary differences associated with the distance from the origin of replication in both the *Shigellae* and *E. coli*; in general this effect is clearest at nucleotide site types where there is a strong distinction of the metric values of the two genera (figs. 5.4.2b, 5.4.3b, 5.5.1e & 5.5.3d). In each case the pattern observed is such that the there is a greater proportion of less deleterious changes closer to the Origin, i.e. increased proportion of SNPs at the 3rd codon position, proportion of SNPs at Q sites, an increase in Transitions over Transversions and a decrease in the proportion of AT enriching SNPs. These patterns are evident as a gradual decrease in the proportion of the less deleterious SNP type along the alignment, with the exception of the +AT/+GC ratio which only shows an origin centred effect within 300-400 genes (fig. 5.5.3d-f).

7.1.4 – *Shigella sonnei* as a special case.

The identification of distinct patterns of evolution in *S. sonnei*, which has only recently adopted the intracellular pathogenic lifestyle of the *Shigellae*, provides strong support for the potential of the methods used in this study to detect evolutionary differences even over very short divergence times.

In cases where there is a clear time-dependent trend, *S. sonnei* shows a pattern of nucleotide changes more consistent with those observed in the *E. coli* than those observed in the other *Shigellae* (figs. 3.6.3a-c, 3.6.4a-b, 4.3.2a-c & 4.3.3a). This is clearest when considering the scatter plots of the residuals to the regression lines for Ti/Tv & +GC+AT at NQ sites (fig. 3.6.5a), where both metric ratios show a clear time dependence.

The simulated equilibrium values for *S. sonnei* are also closer to the *E. coli* than some of the other *Shigellae*, however in this case *S. dysenteriae* also shows a similarly *E. coli* like pattern, which cannot readily be explained but potentially points to hitherto unnoticed *E. coli*-like patterns.

7.1.5 – Overall Conclusions from Results

The time-dependence of purifying selection identified using d*N*/d*S* (Rocha, Maynard Smith et al. 2006) is evident when considering changes at all nucleotide sites or when considering subsets of those sites where there is variation in the selective consequences of the changes.

Additionally this pattern of molecular evolution is detectable via a variety of metrics in addition to dN/dS; the distribution of SNPs among nucleotide site types (Q vs NQ sites or 1st vs 2nd vs 3rd codon position), two metric ratios Ti/Tv & +AT/+GC and the mean metabolic cost of amino acid changes.

The two metric ratios appear to be reflecting slightly different selective pressures, with Ti/Tv showing patterns consistent with the conservation of Amino acid sequences with a time dependant trend towards a greater proportion of the more conservative Transitions (Ti) which is absent from degenerate sites. The +AT/+GC ratio shows patterns consistent with the conservation of features, such as genomic nucleotide content, codon bias and amino acid metabolic cost, as it shows a time dependant trend towards reduced AT enrichment (i.e. increased purging of +AT SNPs) which is evident and largely consistent across all nucleotide site types.

The patterns observed in the various metrics, both when considering the genome as a whole and when examining the genome via a sliding window approach, highlight the *Shigellae* as having a greater proportion of the more deleterious SNPs than *E. coli*, where there is a time dependant trend. This reflects reduced purifying selection in the *Shigellae* which is likely a consequence of their intracellular pathogenic lifestyle, the intracellular mode of replication reducing their effective population size and thus limiting the power of purifying selection to purge deleterious changes. The simulated equilibrium AT content of the *Shigellae* genomes reflects this as they show resulting genome compositions more akin to endosymbionts and obligate intracellular pathogens than do the *E. coli*.

The singular exception to the above is *S. sonnei* which shows a pattern of nucleotide changes more akin to those observed in the *E. coil*. This is likely due to both its relatively

recent adoption of the intracellular lifestyle (~10,000 years ago) and its ability to survive in an environmental host (*Acanthamoeba*) which result in both a lack of sufficient evolutionary time for differences to become apparent and also provide the potential for *S. sonnei* to maintain a much larger effective population size than the other *Shigellae*.

7.2 – Limitations and Further Work

7.2.1 – Sample Size and Taxa Coverage

The sampling of taxa, whilst informative, is incomplete, during the course of the project additional *E. coli* and *Shigella* taxa have been sequenced such that there are now a total of 30 complete genome sequences listed in the CBS Genome Atlas Database (http://www.cbs.dtu.dk/services/GenomeAtlas-3.0/) of which 23 are *E. coli* and 7 *Shigellae*, additionally there is a genome sequence of the closely related outgroup *Escherichia fergusonii*.

This immediately provides a logical expansion of this work, the inclusion of additional taxa would potentially provide more branch lengths or timepoints allowing for a more statistically robust analysis of the trends apparent without obfuscating the signal by including too much 'new' information as the extra taxa will still belong to the same 'species' or group of species as was under study to begin with. The counterpoint to this is that the addition of the extra taxa will consequently reduce the number of SNPs unique to each taxon or branch of the phylogenetic tree, necessitating the use of larger windows for the Around Genome or Along Alignment analysis (Ch5) as that analysis already is subject to some errors associated with windows with low total numbers of SNPs.

An additional expansion along similar lines is to use this approach to examine other groups of taxa which show very close genetic relationships but display very different phenotypes. One such example is that of the genus *Neisseria*, which contains meningitis causing *N. meningitidis*, a major sexually transmitted pathogen *N. gonorrhoeae* and the commensal *N. lactamica*, all of which show a large overlap in gene content (Snyder, Jarvis et al. 2005; Snyder and Saunders 2006). The NCBI Genome database lists 20 *Neisseria* genomes which are currently being sequenced in addition to 6 already completed genomes, which is sufficient for a fairly robust examination of the evolutionary similarities or differences which are evident in the polymorphisms unique to each genome.

7.2.2 – Around Genome or Along Alignment

Due to time constraints only an initial examination of the variation in the patterns of evolution around the genome was performed. However there are a few ways in which this analysis could be improved and expanded upon.

Based upon results observed in Ch4 (& Ch3) it would be logical to restrict the around genome analysis of Ti/Tv to codon position 1 as that is the only codon position which displays a strong evolutionary signal, also the presence of a strong signal across all site types for +AT/+GC suggests that only the analysis of 'All sites' is needed.

Additional analyses which could be performed include a study of the variation in the d*N*/d*S* ratio around the genome and an examination of the time dependence of any given metric at different points around the genome, which would provide a comprehensive insight into the variation of evolutionary forces around the genome.

7.2.3 – Subsets of Genes

The analysis performed solely on the d*N*/d*S* of functional categories of genes could potentially be extended to include the full range of analyses performed in Ch3 & Ch4, allowing a more detailed examination of where the evolutionary signal is coming from, if indeed there is one specific group of genes responsible for the signal. This analysis would also highlight any groups or classes of genes which are undergoing more rapid evolution than the rest of the core genome or are experiencing stronger purifying selection than the rest of the core genome. This would provide more details as to what particular forces are governing the balance of selection and mutation in the genomes examined.

7.2.4 – Problem of Hypermutators

Whilst in the instance of the *E. coli* / *Shigellae* dataset the presence of hypermutators can be excluded, based upon all the genomes having a full complement of functional DNA repair genes, this would need to be verified for any additional taxa included or for other datasets as a hypermutators would resemble a genome experiencing weakened purifying selection. However, by restricting the genomes used to only those which have been fully annotated, or whose repair gene complement is known, and by virtue of the fact that the analysis is centred on 'recent' SNPs, the problem can be largely overcome.

7.2.5 – Simulation of Equilibrium AT Content

Both of the static simulation methods provide a picture of the 'evolutionary heading' of the genomic AT content and in that regard are sufficient for their purpose, however the dynamic simulation method relies heavily upon the assumption that a linear regression to the known points is sufficient to 'evolve' the polymorphism profile.

This has potential for improvement, including the use of either a different regression technique for estimating the trend from the observed points or increasing the number of observed points through the use of a larger dataset, potentially increasing the support and accuracy of the trend used to extrapolate the polymorphism profile. Additionally the estimation of the number of observable SNPs could be improved by including a term in the equation to account for the occurrence of SNP reversions.

This approach could be extended such that the location of a new mutation, in terms of nucleotide site type, could be taken into account with each site having its own set of polymorphism profiles and time dependant trends, allowing for a more detailed and potentially more accurate estimation of the evolutionary 'heading' or 'path' that the genome's AT content is taking.

7.3 – Implications

7.3.1 – d*N*/d*S* – Insufficient as a Measure of Selection

It has been shown previously that the d*N*/d*S* ratio is dependent on divergence time (Rocha, Maynard Smith et al. 2006) and that its ability to correctly distinguish the mode of selection acting between sequences is similarly dependent on how divergent the sequences are; with d*N*/d*S* becoming more insensitive to the value of the selection coefficient between closely related sequences, where the majority of variation is due to segregating polymorphisms rather than fixed substitutions (Kryazhimskiy and Plotkin 2008).

This study has further shown that although the d*N*/d*S* ratio can provide some information as to the relative rates of nucleotide change within and between closely related taxa, the full picture of the evolutionary processes at work is best elucidated via the examination of the patterns of SNPs in the polymorphism profile, summarised in this study by the metric ratios Ti/Tv and +AT/+GC. These provide additional information not readily evident in the d*N*/d*S* ratio, especially where time dependant trends are taken into account, Ti/Tv reflecting patterns associated with rates of change of amino acids, akin to d*N*/d*S* but providing greater insight via the examination of different nucleotide site types, and +GC/+AT which is indicative of trends associated with changes in nucleotide content, independent of Ti/Tv (see comparison of All / NQ / Q sites, Ch 3).

Overall d*N*/d*S* becomes a somewhat blunt instrument for the examination of evolutionary patterns and trends over short timescales, and the methods used in this study either with the chosen metric ratios or with other summaries of the polymorphism profile provide the potential for much more detailed examination of the evolutionary processes at work.

7.3.2 – Inaccuracies of Phylogenetic Reconstruction

In general the construction of a phylogenetic tree requires assumptions, otherwise it becomes impossible to assay which tree is the 'best' and/or the problem is computationally intractable in a sensible amount of time. However there are some assumptions which the results of this study suggest are in error, specifically in the case of reconstructing the trees of closely related species or taxa.

The primary assumption is the lack of 'time' in the evolutionary models specified, in that the parameters of the model, e.g. substitution rates and nucleotide frequencies, are assumed to be static and to apply backwards through evolutionary time, such as in the Generalised Time Reversible (GTR) model. As a result the trees more closely reflect the current observable differences between the taxa rather than inferring the past evolutionary paths of the taxa and where/when they are likely to have diverged.

Specifically in the case of the Maximum Likelihood method of phylogenetic reconstruction, there is a preference to assign observed nucleotide changes to internal nodes, which in turn is based upon the assumption that the vast majority of the nucleotide differences observed are fixed substitutions. However in the case of more closely related taxa a sizeable proportion of the observable differences between taxa represent standing polymorphisms within each taxon or species.

7.3.3 – Identification and Classification of ‘Species’

The current nucleotide based definition of a bacterial species using the cut-off values of 70% for DNA-DNA hybridisation and the approximately equivalent average nucleotide identity of 95% is at best a rough approximation used to conveniently subdivide bacterial taxa into related groupings.

However given the results observed in this study even closely related taxa, in this instance where no two taxa differ at more than 2.65% of their core genome (~5.5% variable sites across the alignment as a whole), it is clear that within the cut-off defined species there are groups of strains/taxa with different lifestyles and consequently experience different patterns of selection which are not evident in their extant genome composition.

The methods used in this study could therefore be applied to groups of closely related taxa, either members of a single formal species or members of species within a species complex, to identify groups of those taxa which are experiencing similar patterns of molecular evolution as evident in their polymorphism profiles. The patterns of evolution within each group could potentially reflect a shared lifestyle or environment; this in turn suggests these taxa as a single ‘ecotype’ and potentially a separate species.

Where this has implications for the application of conventional notions of a ‘species’ is that a species is generally held to share common genetic ancestry, usually visualised as forming a single phylogenetic clade, which in many cases closely related strains will not, as the adoption of the different niche will have come via the acquisition of additional genetic material though horizontal gene transfer, often into different genetic backgrounds. This is notably the case in the *Shigellae* – many of the strains of the four formal species arising independently from several different ancestral *E. coli* lineages via the acquisition of the pINV plasmid.

Consequently at the lowest level individual species cannot be defined solely upon their genetic relatedness; their evolutionary headings and trajectories as well as their ecological niche must also be taken into account in order to build a complete picture of their relationships.

References

Akashi, H. and T. Gojobori (2002). "Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis." Proc Natl Acad Sci U S A **99**(6): 3695-3700.

Andersson, S. G., C. Alsmark, et al. (2002). "Comparative genomics of microbial pathogens and symbionts." Bioinformatics **18 Suppl 2**: S17.

Andersson, S. G. and C. G. Kurland (1998). "Reductive evolution of resident genomes." Trends Microbiol **6**(7): 263-268.

Andersson, S. G., A. Zomorodipour, et al. (1998). "The genome sequence of Rickettsia prowazekii and the origin of mitochondria." Nature **396**(6707): 133-140.

Avery, O. T., C. M. MacLeod, et al. (1995). "Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944." Mol Med **1**(4): 344-365.

Banerjee, T., S. K. Gupta, et al. (2005). "Role of mutational bias and natural selection on genome-wide nucleotide bias in prokaryotic organisms." Biosystems **81**(1): 11-18.

Beadle, G. W. and E. L. Tatum (1941). "Genetic Control of Biochemical Reactions in Neurospora." Proc Natl Acad Sci U S A **27**(11): 499-506.

Bergey, D. H. and S. o. A. Bacteriologists (1930). Bergey's Manual of Determinative Bacteriology. London, Bailliere, Tindall and Cox.

Bipatnath, M., P. P. Dennis, et al. (1998). "Initiation and velocity of chromosome replication in Escherichia coli B/r and K-12." J Bacteriol **180**(2): 265-273.

Birky, C. W., Jr., T. Maruyama, et al. (1983). "An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results." Genetics **103**(3): 513-527.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1474.

Bopp, C., F. Brenner, et al. (2003). *Escherichia*, *Shigella*, and *Salmonella*. Manual of Clinical Microbiology. E. J. B. P. R. Murray, J. H. Jorgensen, M. A. Pfaller, and R. H. Yolken. Washington, D.C., ASM Press. **1:** 654-671.

Butler, T., P. Speelman, et al. (1986). "Colonic dysfunction during shigellosis." J Infect Dis **154**(5): 817-824.

Campo, N., M. J. Dias, et al. (2004). "Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions." Mol Microbiol **51**(2): 511-522.

Chamary, J. V. and L. D. Hurst (2005). "Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals." Genome Biol **6**(9): R75.

Chargaff, E. (1950). "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation." Experientia **6**(6): 201-209.

Chen, S. L., C. S. Hung, et al. (2006). "Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach." Proc Natl Acad Sci U S A **103**(15): 5977-5982.

Cole, S. T., K. Eiglmeier, et al. (2001). "Massive gene decay in the leprosy bacillus." Nature **409**(6823): 1007-1011.

Collins, D. W. and T. H. Jukes (1994). "Rates of transition and transversion in coding sequences since the human-rodent divergence." Genomics **20**(3): 386-396.

Cooper, S. and C. E. Helmstetter (1968). "Chromosome replication and the division cycle of Escherichia coli B/r." J Mol Biol **31**(3): 519-540.

Couturier, E. and E. P. Rocha (2006). "Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes." Mol Microbiol **59**(5): 1506-1518.

Craig, C. L. and R. S. Weber (1998). "Selection costs of amino acid substitutions in ColE1 and ColIa gene clusters harbored by Escherichia coli." Mol Biol Evol **15**(6): 774-776.

Dale, C. and N. A. Moran (2006). "Molecular interactions between bacterial symbionts and their hosts." Cell **126**(3): 453-465.

Dancey, G. F. and B. M. Shapiro (1976). "The NADH dehydrogenase of the respiratory chain of Escherichia coli. II. Kinetics of the purified enzyme and the effects of antibodies elicited against it on membrane-bound and free enzyme." J Biol Chem **251**(19): 5921-5928.

Daubin, V. and G. Perriere (2003). "G+C3 structuring along the genome: a common feature in prokaryotes." Mol Biol Evol **20**(4): 471-483.

Denver, D. R., K. Morris, et al. (2004). "High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome." Nature **430**(7000): 679-682.

Dom, P., F. Haesebrouck, et al. (1994). "In vivo association of Actinobacillus pleuropneumoniae serotype 2 with the respiratory epithelium of pigs." Infect Immun **62**(4): 1262-1267.

Donohue-Rolfe, A., D. W. Acheson, et al. (1991). "Shiga toxin: purification, structure, and function." Rev Infect Dis **13 Suppl 4**: S293-297.

Dumler, J. S., J. E. Madigan, et al. (2007). "Ehrlichioses in humans: epidemiology, clinical presentation, diagnosis, and treatment." Clin Infect Dis **45 Suppl 1**: S45-51.

Duncan, B. K. and J. H. Miller (1980). "Mutagenic deamination of cytosine residues in DNA." Nature **287**(5782): 560-561.

DuPont, H. L., M. M. Levine, et al. (1989). "Inoculum size in shigellosis and implications for expected mode of transmission." J Infect Dis **159**(6): 1126-1128.

Dutta, S., K. Rajendran, et al. (2002). "Shifting serotypes, plasmid profile analysis and antimicrobial resistance pattern of shigellae strains isolated from Kolkata, India during 1995-2000." Epidemiol Infect **129**(2): 235-243.

Enright, M. C. and B. G. Spratt (1999). "Multilocus sequence typing." Trends Microbiol **7**(12): 482-487.

Escobar-Paramo, P., S. Ghosh, et al. (2005). "Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic archaeon Pyrococcus." Mol Biol Evol **22**(11): 2297-2303.

Feil, E. J., J. E. Cooper, et al. (2003). "How clonal is Staphylococcus aureus?" J Bacteriol **185**(11): 3307-3316.

Felsenstein, J. (1989). "PHYLIP - Phylogeny Inference Package (Version 3.2)." Cladistics **5**(2): 164-166.

Flexner, S. and L. F. Barker (1900). "The Prevalent Diseases in the Philippines." Science **11**(275): 521-528.

Foerstner, K. U., C. von Mering, et al. (2005). "Environments shape the nucleotide composition of genomes." EMBO Rep **6**(12): 1208-1213.

Francis, E. (1921). "Tularemia. I. The occurrence of tularemia in nature as a disease of man." Public Health Reports **36**: 1731-1753.

Franklin, R. E. and R. G. Gosling (1953). "Molecular configuration in sodium thymonucleate." Nature **171**(4356): 740-741.

Freeland, S. J. and L. D. Hurst (1998). "The genetic code is one in a million." J Mol Evol **47**(3): 238-248.

Fremaux, B., C. Prigent-Combaret, et al. (2008). "Long-term survival of Shiga toxin-producing Escherichia coli in cattle effluents and environment: an updated review." Vet Microbiol **132**(1-2): 1-18.

Galtier, N. and J. R. Lobry (1997). "Relationships Between Genomic G+C Content, RNA Secondary Structures, and Optimal Growth Temperature in Prokaryotes." Journal of Molecular Evolution **44**(6): 632-636.

Goldberg, M. B. (2001). "Actin-based motility of intracellular microbial pathogens." Microbiol Mol Biol Rev **65**(4): 595-626, table of contents.

Gupta, A., C. S. Polyak, et al. (2004). "Laboratory-confirmed shigellosis in the United States, 1989-2002: epidemiologic trends and patterns." Clin Infect Dis **38**(10): 1372-1377.

Gupta, S. K., S. Majumdar, et al. (2000). "Studies on the Relationships between the Synonymous Codon Usage and Protein Secondary Structural Units." Biochemical and Biophysical Research Communications **269**(3): 692-696.

Gutacker, M. M., J. C. Smoot, et al. (2002). "Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains." Genetics **162**(4): 1533-1543.

Haig, D. and L. D. Hurst (1991). "A quantitative measure of error minimization in the genetic code." J Mol Evol **33**(5): 412-417.

Haldane, J. B. S. (1957). "Cost of Natural Selection." Journal of Genetics **55**: 511-524.

Halliday, J. A. and B. W. Glickman (1991). "Mechanisms of spontaneous mutation in DNA repair-proficient Escherichia coli." Mutat Res **250**(1-2): 55-71.

Hallin, P. F. and D. W. Ussery (2004). "CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data." Bioinformatics **20**(18): 3682-3686.

Hamady, M., M. D. Betterton, et al. (2006). "Using the nucleotide substitution rate matrix to detect horizontal gene transfer." BMC Bioinformatics **7**: 476.

Hardy, G. H. (1908). "Mendelian Proportions in a Mixed Population." Science **28**(706): 49-50.

Hartl, D. L., E. N. Moriyama, et al. (1994). "Selection intensity for codon bias." Genetics **138**(1): 227-234.

Hedges, S. R., W. W. Agace, et al. (1995). "Epithelial cytokine responses and mucosal cytokine networks." Trends Microbiol **3**(7): 266-270.

Heizer, E. M., Jr., D. W. Raiford, et al. (2006). "Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis." Mol Biol Evol **23**(9): 1670-1680.

Hershberg, R., H. Tang, et al. (2007). "Reduced selection leads to accelerated gene loss in Shigella." Genome Biol **8**(8): R164.

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.

Hiller, N. L., B. Janto, et al. (2007). "Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome." J Bacteriol **189**(22): 8186-8195.

Ho, S. Y. and G. Larson (2006). "Molecular clocks: when times are a-changin'." Trends Genet **22**(2): 79-83.

Ho, S. Y., M. J. Phillips, et al. (2005). "Time dependency of molecular rate estimates and systematic overestimation of recent divergence times." Mol Biol Evol **22**(7): 1561-1568.

Hogg, J. S., F. Z. Hu, et al. (2007). "Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains." Genome Biol **8**(6): R103.

Hormaeche, E. and C. A. Peluffo (1959). "Laboratory diagnosis of Shigella and Salmonella infections." Bull World Health Organ **21**: 247-277.

Hughes, A. L., R. Friedman, et al. (2008). "Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes." Mol Biol Evol **25**(10): 2199-2209.

Hughes, A. L., B. Packer, et al. (2003). "Widespread purifying selection at polymorphic sites in human protein-coding loci." Proc Natl Acad Sci U S A **100**(26): 15754-15757.

Hurst, L. D., E. J. Feil, et al. (2006). "Protein evolution: causes of trends in amino-acid gain and loss." Nature **442**(7105): E11-12; discussion E12.

Ishii, S., W. B. Ksoll, et al. (2006). "Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds." Appl Environ Microbiol **72**(1): 612-621.

Jeong, H. J., E. S. Jang, et al. (2007). "Acanthamoeba: could it be an environmental host of Shigella?" Exp Parasitol **115**(2): 181-186.

Jin, Q., Z. Yuan, et al. (2002). "Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157." Nucleic Acids Res **30**(20): 4432-4441.

Johnson, K. P. and J. Seger (2001). "Elevated rates of nonsynonymous substitution in island birds." Mol Biol Evol **18**(5): 874-881.

Jordan, I. K., F. A. Kondrashov, et al. (2005). "A universal trend of amino acid gain and loss in protein evolution." Nature **433**(7026): 633-638.

Jordan, I. K., I. B. Rogozin, et al. (2002). "Microevolutionary genomics of bacteria." Theor Popul Biol **61**(4): 435-447.

Karaolis, D. K., R. Lan, et al. (1994). "Sequence variation in Shigella sonnei (Sonnei), a pathogenic clone of Escherichia coli, over four continents and 41 years." J Clin Microbiol **32**(3): 796-802.

Keusch, G. T. and M. L. Bennish (1998). Shigellosis. Bacterial Infections of Humans: Epidemiology & Control. A. S. Evans and P. S. Brachman. New York, Plenum Publishing Co.**:** 631-656.

Kimura, M. (1968). "Evolutionary rate at the molecular level." Nature **217**(5129): 624-626.

Kimura, M. (1991). "Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics." Proc Natl Acad Sci U S A **88**(14): 5969-5973.

King, J. L. and T. H. Jukes (1969). "Non-Darwinian evolution." Science **164**(881): 788-798.

Knight, R. D., S. J. Freeland, et al. (2001). "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes." Genome Biol **2**(4): RESEARCH0010.

Kotloff, K. L., J. P. Winickoff, et al. (1999). "Global burden of Shigella infections: implications for vaccine development and implementation of control strategies." Bull World Health Organ **77**(8): 651-666.

Kryazhimskiy, S. and J. B. Plotkin (2008). "The population genetics of dN/dS." PLoS Genet **4**(12): e1000304.

Kunz, B. A. and S. E. Kohalmi (1991). "Modulation of mutagenesis by deoxyribonucleotide levels." Annu Rev Genet **25**: 339-359.

Kurland, C. G. (1991). "Codon bias and gene expression." FEBS Lett **285**(2): 165-169.

Lan, R., M. C. Alles, et al. (2004). "Molecular evolutionary relationships of enteroinvasive Escherichia coli and Shigella spp." Infect Immun **72**(9): 5080-5088.

Lau, M. M. and S. C. Ingham (2001). "Survival of faecal indicator bacteria in bovine manure incorporated into soil." Lett Appl Microbiol **33**(2): 131-136.

Leclerc, H., L. Schwartzbrod, et al. (2002). "Microbial agents associated with waterborne diseases." Crit Rev Microbiol **28**(4): 371-409.

Levine, O. S. and M. M. Levine (1991). "Houseflies (Musca domestica) as mechanical vectors of shigellosis." Rev Infect Dis **13**(4): 688-696.

Lind, P. A. and D. I. Andersson (2008). "Whole-genome mutational biases in bacteria." Proc Natl Acad Sci U S A **105**(46): 17878-17883.

Majumdar, S., S. K. Gupta, et al. (1999). "Compositional Correlation Studies among the Three Different Codon Positions in 12 Bacterial Genomes." Biochemical and Biophysical Research Communications **266**(1): 66-71.

Makino, K., K. Yokoyama, et al. (1999). "Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic Escherichia coli O157:H7 derived from the Sakai outbreak." Genes Genet Syst **74**(5): 227-239.

Marais, G. A., A. Calteau, et al. (2008). "Mutation rate and genome reduction in endosymbiotic and free-living bacteria." Genetica **134**(2): 205-210.

Maurelli, A. T., R. E. Fernandez, et al. (1998). ""Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli." Proc Natl Acad Sci U S A **95**(7): 3943-3948.

McDonald, J. H. (2006). "Apparent trends of amino Acid gain and loss in protein evolution due to nearly neutral variation." Mol Biol Evol **23**(2): 240-244.

McOrist, S., C. J. Gebhart, et al. (1995). "Characterization of Lawsonia intracellularis gen. nov., sp. nov., the obligately intracellular bacterium of porcine proliferative enteropathy." Int J Syst Bacteriol **45**(4): 820-825.

Medini, D., D. Serruto, et al. (2008). "Microbiology in the post-genomic era." Nat Rev Microbiol **6**(6): 419-430.

Menard, R., C. Dehio, et al. (1996). "Bacterial entry into epithelial cells: the paradigm of Shigella." Trends Microbiol **4**(6): 220-226.

Michaels, M. L. and J. H. Miller (1992). "The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine)." J Bacteriol **174**(20): 6321-6325.

Minitab (2006). Minitab Statistical Software, Release 15 for Windows. State College, Pennsylvania, Minitab® is a registered trademark of Minitab Inc.

Mira, A. and N. A. Moran (2002). "Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria." Microb Ecol **44**(2): 137-143.

Mira, A. and H. Ochman (2002). "Gene location and bacterial sequence divergence." Mol Biol Evol **19**(8): 1350-1358.

Mirkin, E. V. and S. M. Mirkin (2005). "Mechanisms of transcription-replication collisions in bacteria." Mol Cell Biol **25**(3): 888-895.

Moran, N. A. (2002). "Microbial minimalism: genome reduction in bacterial pathogens." Cell **108**(5): 583-586.

Moran, N. A., H. J. McLaughlin, et al. (2009). "The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria." Science **323**(5912): 379-382.

Moran, N. A. and G. R. Plague (2004). "Genomic changes following host restriction in bacteria." Curr Opin Genet Dev **14**(6): 627-633.

Muto, A. and S. Osawa (1987). "The guanine and cytosine content of genomic DNA and bacterial evolution." Proc Natl Acad Sci U S A **84**(1): 166-169.

Naya, H., H. Romero, et al. (2002). "Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes." J Mol Evol **55**(3): 260-264.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-426.

Neutra, M. R., A. Frey, et al. (1996). "Epithelial M cells: gateways for mucosal infection and immunization." Cell **86**(3): 345-348.

Niki, H. and S. Hiraga (1998). "Polar localization of the replication origin and terminus in Escherichia coli nucleoids during chromosome partitioning." Genes & Development **12**(7): 1036-1045.

Niyogi, S. K. (2005). "Shigellosis." J Microbiol **43**(2): 133-143.

Ochman, H. (2003). "Neutral mutations and neutral substitutions in bacterial genomes." Mol Biol Evol **20**(12): 2091-2096.

Ochman, H., T. S. Whittam, et al. (1983). "Enzyme polymorphism and genetic population structure in Escherichia coli and Shigella." J Gen Microbiol **129**(9): 2715-2726.

Ohta, T. (1973). "Slightly deleterious mutant substitutions in evolution." Nature **246**(5428): 96-98.

Oliver, J. D. (2005). "Wound infections caused by Vibrio vulnificus and other marine bacteria." Epidemiol Infect **133**(3): 383-391.

Pal, S. C. (1984). "Epidemic bacillary dysentery in West Bengal, India, 1984." Lancet **1**(8392): 1462.

Parkhill, J., M. Sebaihia, et al. (2003). "Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica." Nat Genet **35**(1): 32-40.

Parkhill, J., B. W. Wren, et al. (2001). "Genome sequence of Yersinia pestis, the causative agent of plague." Nature **413**(6855): 523-527.

Perez-Brocal, V., R. Gil, et al. (2006). "A small microbial genome: the end of a long symbiotic relationship?" Science **314**(5797): 312-313.

Perry, R. D. and J. D. Fetherston (1997). "Yersinia pestis--etiologic agent of plague." Clin Microbiol Rev **10**(1): 35-66.

Petersen, L., J. P. Bollback, et al. (2007). "Genes under positive selection in Escherichia coli." Genome Res **17**(9): 1336-1343.

Pupo, G. M., D. K. Karaolis, et al. (1997). "Evolutionary relationships among pathogenic and nonpathogenic Escherichia coli strains inferred from multilocus enzyme electrophoresis and mdh sequence studies." Infect Immun **65**(7): 2685-2692.

Pupo, G. M., R. Lan, et al. (2000). "Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics." Proc Natl Acad Sci U S A **97**(19): 10567-10572.

Ragan, M. A. (2001). "Detection of lateral gene transfer among microbial genomes." Curr Opin Genet Dev **11**(6): 620-626.

Read, T. D., S. L. Salzberg, et al. (2002). "Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis." Science **296**(5575): 2028-2033.

Rocha, E. P. (2004). "The replication-related organization of bacterial genomes." Microbiology **150**(Pt 6): 1609-1627.

Rocha, E. P. (2008). "The organization of the bacterial genome." Annu Rev Genet **42**: 211-233.

Rocha, E. P. and A. Danchin (2002). "Base composition bias might result from competition for metabolic resources." Trends Genet **18**(6): 291-294.

Rocha, E. P. and A. Danchin (2003). "Gene essentiality determines chromosome organisation in bacteria." Nucleic Acids Res **31**(22): 6570-6577.

Rocha, E. P., I. Matic, et al. (2002). "Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions?" Nucleic Acids Res **30**(9): 1886-1894.

Rocha, E. P., J. Maynard Smith, et al. (2006). "Comparisons of dN/dS are time dependent for closely related bacterial genomes." J Theor Biol **239**(2): 226-235.

Rocha, E. P. C., E. Cornet, et al. (2005). "Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems." PLoS Genetics **1**(2): e15.

Rolland, K., N. Lambert-Zechovsky, et al. (1998). "Shigella and enteroinvasive Escherichia coli strains are derived from distinct ancestral strains of E. coli." Microbiology **144 (Pt 9)**: 2667-2672.

Saeed, A., H. Abd, et al. (2008). "Acanthamoeba castellanii an environmental host for Shigella dysenteriae and Shigella sonnei." Arch Microbiol.

Sallstrom, B. and S. G. Andersson (2005). "Genome reduction in the alpha-Proteobacteria." Curr Opin Microbiol **8**(5): 579-585.

Sansonetti, P. J., H. d'Hauteville, et al. (1982). "Plasmid-mediated invasiveness of "Shigella-like" Escherichia coli." Ann Microbiol (Paris) **133**(3): 351-355.

Sansonetti, P. J. and A. Phalipon (1999). "M cells as ports of entry for enteroinvasive pathogens: mechanisms of interaction, consequences for the disease process." Semin Immunol **11**(3): 193-203.

Schaaper, R. M. and R. L. Dunn (1991). "Spontaneous mutation in the Escherichia coli lacI gene." Genetics **129**(2): 317-326.

Sharp, P. M., D. C. Shields, et al. (1989). "Chromosomal location and evolutionary rate variation in enterobacterial genes." Science **246**(4931): 808-810.

Shepherd, J. G., L. Wang, et al. (2000). "Comparison of O-antigen gene clusters of Escherichia coli (Shigella) sonnei and Plesiomonas shigelloides O17: sonnei gained its current plasmid-borne O-antigen genes from P. shigelloides in a recent event." Infect Immun **68**(10): 6056-6061.

Shiga, K. (1898). "Ueber den Dysenterie bacillus (*Bacillus dysenteriae*)." Zentralbl Bakteriol Parasitenkd Abt I Org. **24**: 817-824.

Shiga, K. (1906). "Observation on the epidemiology of dysentery in Japan." Philippine J. of Sci. **1**: 485-500.

Silva, R. M., M. R. Toledo, et al. (1982). "Correlation of invasiveness with plasmid in enteroinvasive strains of Escherichia coli." J Infect Dis **146**(5): 706.

Singer, G. A. and D. A. Hickey (2000). "Nucleotide bias causes a genomewide bias in the amino acid composition of proteins." Mol Biol Evol **17**(11): 1581-1588.

Snyder, L. A., S. A. Jarvis, et al. (2005). "Complete and variant forms of the 'gonococcal genetic island' in Neisseria meningitidis." Microbiology **151**(Pt 12): 4005-4013.

Snyder, L. A. and N. J. Saunders (2006). "The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica, including those designated as 'virulence genes'." BMC Genomics **7**: 128.

Sousa, C., V. de Lorenzo, et al. (1997). "Modulation of gene expression through chromosomal positioning in Escherichia coli." Microbiology **143 (Pt 6)**: 2071-2078.

Sultana, I., R. M. Mizanur, et al. (2002). "Survivality and Virulence of *Shigella sonnei* and *Shigella boydii* in Different Physico-Chemical Stress Conditions." Journal of Biological Sciences **2**(3): 196-201.

Swire, J. (2007). "Selection on synthesis cost affects interprotein amino acid usage in all three domains of life." J Mol Evol **64**(5): 558-571.

Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." Mol Biol Evol **24**(8): 1596-1599.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.

Touchon, M., C. Hoede, et al. (2009). "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths." PLoS Genet **5**(1): e1000344.

Ussery, D., T. S. Larsen, et al. (2001). "Genome organisation and chromatin structure in Escherichia coli." Biochimie **83**(2): 201-212.

van Passel, M. W., P. R. Marri, et al. (2008). "The emergence and fate of horizontally acquired genes in Escherichia coli." PLoS Comput Biol **4**(4): e1000059.

van Pelt, W., M. A. de Wit, et al. (2003). "Laboratory surveillance of bacterial gastroenteric pathogens in The Netherlands, 1991-2001." Epidemiol Infect **130**(3): 431-441.

Vladimirov, N., V. Likhoshvai, et al. (2007). "Correlation of codon biases and potential secondary structures with mRNA translation efficiency in unicellular organisms." Molecular Biology **41**(5): 843-850.

von Seidlein, L., D. R. Kim, et al. (2006). "A Multicentre Study of Shigella Diarrhoea in Six Asian Countries: Disease Burden, Clinical Manifestations, and Microbiology." PLoS Medicine **3**(9): e353.

Wang, H. C., E. Susko, et al. (2006). "On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors." Biochem Biophys Res Commun **342**(3): 681-684.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.

Welch, R. A., V. Burland, et al. (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli." Proc Natl Acad Sci U S A **99**(26): 17020-17024.

Wernegreen, J. J. (2002). "Genome evolution in bacterial endosymbionts of insects." Nat Rev Genet **3**(11): 850-861.

Wernegreen, J. J. and D. J. Funk (2004). "Mutation exposed: a neutral explanation for extreme base composition of an endosymbiont genome." J Mol Evol **59**(6): 849-858.

Wernegreen, J. J. and N. A. Moran (1999). "Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes." Mol Biol Evol **16**(1): 83-97.

Wilkins, M. H., A. R. Stokes, et al. (1953). "Molecular structure of deoxypentose nucleic acids." Nature **171**(4356): 738-740.

Wirth, T., D. Falush, et al. (2006). "Sex and virulence in Escherichia coli: an evolutionary perspective." Mol Microbiol **60**(5): 1136-1151.

Woolfit, M. and L. Bromham (2003). "Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes." Mol Biol Evol **20**(9): 1545-1555.

Woolfit, M. and L. Bromham (2005). "Population size and molecular evolution on islands." Proc Biol Sci **272**(1578): 2277-2282.

Wu, M., L. V. Sun, et al. (2004). "Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements." PLoS Biol **2**(3): E69.

Yang, F., J. Yang, et al. (2005). "Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery." Nucleic Acids Res **33**(19): 6445-6458.

Yang, J., H. Nie, et al. (2007). "Revisiting the Molecular Evolutionary History of Shigella spp." J Mol Evol **64**(1): 71-79.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.

Zhang, J. (2000). "Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes." J Mol Evol **50**(1): 56-68.

Appendices

Appendix I "The Rise and Fall of Deleterious Mutation"

Appendix II "The temporal dynamics of slightly deleterious mutations in Escherichia coli and Shigella spp."

Appendix IIIa Description of Perl and Shell scripts used

Appendix IIIb CD-ROM containing the Perl and Shell scripts used

Appendix I – “The Rise and Fall of Deleterious Mutation”

Balbi, K. J. Feil, E. J.

Published in *Research in Microbiology* in 2007

“It is well established that selection is less efficient in small populations than in large ones. Here we review the impact of this effect by considering the gradual selective purging of deleterious mutation over time. We outline an approach to explore the dynamics of this process, and highlight its profound implications.”

Appendix II – “The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella* spp”

Balbi, K. J. Rocha, E. P. Feil, E. J.

Published in *Molecular Biology and Evolution* in 2009

“The Shigella are recently emerged clones of Escherichia coli, which have independently adopted an intracellular pathogenic lifestyle. We examined the molecular evolutionary consequences of this niche specialization by comparing the normalized, directional frequency profiles of unique polymorphisms within 2,098 orthologues representing the intersection of five E. coli and four Shigella genomes. We note a surfeit of AT-enriching changes (GC-->AT), transversions, and nonsynonymous changes in the Shigella genomes. By examining these differences within a temporal framework, we conclude that our results are consistent with relaxed or inefficient selection in Shigella owing to a reduced effective population size. Alternative interpretations, and the interesting exception of Shigella sonnei, are discussed. Finally, this analysis lends support to the view that nucleotide composition typically does not lie at mutational equilibrium but that selection plays a role in maintaining a higher GC content than would result solely from mutation bias. This argument sheds light on the enrichment of adenine and thymine in the genomes of bacterial endosymbionts where purifying selection is very weak.”

Appendix IIIa – Description of Scripts

Actual script files are available on the enclosed CD (Appendix IIIb). However the name of the script files, a brief outline of their syntax and an overview of their function is listed below.

SNP Ratio Pipeline

exclude-count-summary.pipe.sh <Alignment File> <Dataset Code> <Number of Taxa>

Requires two additional files both starting with the dataset code – a .id file containing a list of the taxa and a .ex file containing a list of the taxon exclusions to be performed.

The script sequentially performs each taxon exclusion and then using the *subst* and *qsubst* programs from Eduardo Rocha identifies all of the SNPs unique to each branch at all sites, each codon position and Q sites & summarises these into a single file for each site type using *subs_tot_capture.pl* (below).

Next base counts are performed at each site type and summarised to a single file for each site type using *count.pl* / *qcount.pl* / *codcount.pl* (below)

The SNPs and Base Counts at NQ sites are then determined using the All sites and Q site counts via *subs.diff.pl* and *count.diff.pl* (below)

Finally for each site type the SNPs are normalised and the Ti/Tv & + AT/+GC ratios are calculated via *ratio.calc.pl* (below)

subs_tot_capture.pl <output file of *subst/qsubst*> <Taxon> <Excluded Taxa>

Simply reads in the summary line at the end of the output from *subst/qsubst*, reformats and outputs to the display, is redirected to append to a file in the pipeline script via the ">> output.file" syntax

count.pl/qcount.pl/codcount.pl <alignment file> <output file>

Reads through the sequence data in the alignment file, counts the nucleotide bases and outputs the results to the specified output file

subs.diff.pl / count.diff.pl <All Site data file> <Q Site data file> <Output File>

Reads in the All site and Q site data and calculates the corresponding values for the NQ sites.

ratio.calc.pl <SNP counts> <Base counts> <Output File>

Takes the summarised files for a given site type, calculates the normalised polymorphism profiles and SNP ratio values for each taxon from these values. Outputting the results to the specified Output File.

Bootstrap Analysis

bs.ratio-95ci.pl <SNP & Base Counts File> <Ratio Config File> <Number of replicates>

The first input file specifies one taxon per line with the first 12 elements being the SNP counts (in alphabetical order) the next 4 being the base counts (also alphabetically) and the final element is the taxon name.

The second input file specifies the ratios to calculate the 95% confidence intervals for. Each line specifying (comma delimited) the ratio title, the numerator values, the denominator values. The values themselves being hyphen delimited (e.g. test_ratio,2-4-6-8,1-3-5-7-9)

The SNP counts are resampled with replacement for the number of replicates specified. For each replicate the SNP counts are normalised and the specified ratio values are calculated. Once all replicates are complete the 95% confidence intervals are calculated using the replicate ratio values.

bs.ratio-compare.pl <Reference File> <Test File> <Output File> <Number of Analyses> <Number of Replicates>

Each of the reference and test files comprises of 1 column per data point in with the first 12 rows containing the SNP types in alphabetical order, followed by the total number of SNPs, and then the base counts in alphabetical order.

Each pair of columns (one in test file, one in reference file) comprises as single analysis. The reference data is resampled such that it possesses the same total number of SNPs as the test data, normalised and the Ti/Tv and +AT/+GC ratio calculated. This is repeated for the specified number of replicates, after which the ratio values for the test dataset alone are calculated and compared to the resampled values.

The number of replicates greater than and less than and equal to the test data are counted. Additionally the 5^{th} and 95^{th} percentiles of the replicates are calculated and the test data point value is also represented as a percentile of the replicates.

SNP reversal for d*N*/d*S* within branch Analysis

unsubs.pl <Input FASTA File> <Output FASTA File>

For each of the sequences in the input FASTA file, the type and location of each of the unique SNPs is identified and stored, once complete for all taxa the identified SNPs are systematically reversed and the reverted sequences saved to the Output FASTA file.

Amino Acid Substitution Analysis

AASubs.batch.sh <Alignment File (Nucleotide)> <Dataset Code> <Number of Taxa>

Requires two additional files both starting with the dataset code – a .id file containing a list of the taxa and a .ex file containing a list of the taxon exclusions to be performed.

The script (as with the nucleotide script) sequentially performs each taxon exclusion and then analyses the dataset using the *dna2aasubst* program by Eduardo Rocha which identifies all of the Amino acid changes unique to each taxon based upon the nucleotide sequence data.

This data is then summarised, normalised and the mean metabolic cost per AA change calculated by the script *aa.summary.pl* (below)

aa.summary.pl <Dataset Code> < Exclusion List> <*dna2aasubst* Output Folder>

Summarises all the individual analyses, normalises the data and calculates the mean metabolic cost per amino acid change, saving the output to:

"<Dataset Code>.MeanCost.Results.csv"

TBS Internal Branch Ratio Calculation

ratio.calc.int-branch.pl <Config File> <Taxon ID List> <Dataset Code> <Site Code>

Requires that the script be executed in the same folder as the results from the SNP Ratio Pipeline results are located. The config file contains one line per internal branch as follows:

'Internal Branch ID',Taxon,ExSPP+ExSPP+...,DelSPP+DelSPP+...,SupSPP+SupSPP+...

Where ExSPP = Excluded Taxon, DelSPP = Deleted Taxon and SupSPP = Taxon Supported by the internal branch.

Using the information from the configuration file, each of the internal branch polymorphism profiles is calculated as detailed in the materials and methods section (2.8.1). The results are output directly to the display, but can be redirected to a file via the use of the "> 'output.file'" syntax.

Sliding Window Base Counts

slwin.counts.pl <Input Data> <Output File> <Window Size> <Window Step>

Window size and step are in Number of Genes.

The script iterates through the input, each line containing the nucleotide counts of a single gene, summarising using window sizes and steps specified and saving the results to the Output File.

Sliding Window SNP Counts

slwin.subs.pl <Input Data> <Output File> <Window Size> <Window Step>

Window size and step are in Number of Genes.

The script iterates through the input, which is the output files from the *subst* program when run on an unconcatenated set of sequence alignments, summarising using window sizes and steps specified and saving the results to the Output File.

SNP Density Analysis

SNP.density.pl <Input Alignment> <Output File> <Window Size> <Window Step>

Window size and step are in Kbp.

The script steps along the alignment, using the window sizes and steps specified, identifying and counting the number of SNPs in a given window. The results of this are saved to the Output File.

Stochastic Equilibrium AT Content Simulation

at.sim.static-random.pl <Input File> <Run Length> <Report Rate> <#Simulations>

Input file contains the SNP counts and nucleotide composition of the taxa to be analysed, each row a separate dataset (with the first row containing dataset names).

The script runs the simulation as specified in the Materials and Methods (2.11.1), using the parameters supplied. Output is to the display

Static Matrix Equilibrium AT Content Simulation

at.sim.static-matrix.pl <Input File> <Output File> <Report Rate> <#Iterations>

Input file is the same layout as above except data is the normalised matrix of base interchanges such that the sum of all the changes originating from a given base is equal to one.

The script runs the simulation as specified in the Materials and Methods (2.11.2), using the parameters supplied. Output is saved in the specified output file.

Dynamic Matrix Equilibrium AT Content Simulation

at.sim.dynamic.pl <Input File> <Output File> <Report Rate> <#Iterations> <Output Style>

The additional factor of output style determines the level of detail saved to the output file.

Instead of the input rows containing separate taxa for analysis the first row is the base counts of the genome, and the subsequent rows are the normalised SNP matrix for each of the timepoints for the genome, prefixed by the Log Divergence Time associated with that matrix.

The script runs the simulation as specified in the Materials and Methods (2.11.3), using the parameters supplied. Output is saved in the specified output file.

SNP Matrix Counting

subs.matrix.pl <Alignment File> <Output File>

Identifies SNPs and conserved bases outputting the results in alphabetical order (AA,AC,AG,AT,CA,CC etc) to the output file.

Alignment Segment Isolation (for Bayesian Analysis)

fasta_equal-distrib_chunks.pl / fasta_random_chunks.pl <Alignment File> <Number of Chunks> <Chunk Size>

Chunk size is in Kbp. Both scripts read in the sequence data and isolate alignment chunks of the specified size writing the data to numbered individual fasta files.